

Subjectivity, objectivity and the insights they bring

Discussions about art and science can be frustrating but occasionally stimulating. The Editor explains why this issue includes the first of a series of artistic and art-historical contributions to *Nature*.

Only rarely do the interests of art and of science come into direct conflict — although, as it happens, this week provides three examples. The balance of art and science within the Musée de l'Homme in Paris is now being resolved after heated debate (see page 217). Britain's Labour government's National Endowment for Science, Technology and the Arts, although potentially beneficial for both, nevertheless promises to set them in competition for funds (see page 222). And the artist Eduardo Paolozzi has been at the centre of comments that his newly unveiled sculpture of Newton, in imitating William Blake's allegedly snide portrait, could be seen as casting science in a bad light (see page 221).

Such controversies are untypical. Far more frequent, I believe, are attempts to see both enterprises as positive influences on each other, and to that end, much has been written comparing art and science and attempting to blur the boundaries between the two — with sceptics riposting that such discussions are ultimately of little value. It is indeed surely true that, for some artists and some scientists, the moment of creation can be of similar character — an inner reevaluation of what they have perceived, leading on the one hand to a personal and subjective rendition, and on the other to a logical train of thought and new understanding that is quickly incorporated into an impersonal scientific edifice. It is not clear, however, where all of that leads us.

Furthermore, science has influenced art — as anybody who has taken an interest in artists' understanding and use of colour and perspective throughout history knows — whereas examples of art influencing science are thin on the ground. And both have been more profoundly influenced by technology. The temptation to link the two is particularly strong when considering the turn of the century, when relativity and quantum mechanics transformed scientists' perceptions. At the same time, artists and composers were undertaking a similarly profound re-evaluation of their 'languages', resulting, for example, in expressionism, cubism and atonality. That coincidence of revolutions is tantalizing, but evidence that there was in fact significant cross-fertilization remains elusive.

When Antonia Payne of the Ruskin School of Drawing and Fine Art invited *Nature* to be involved in a project linking us with artists — sponsored by the Arts Council of England — I was somewhat sceptical. Much of the discussion about art and science is of marginal importance and largely irrelevant to practitioners of both, while most people turn to each for very different things. So I asked to look at the works of candidate artists before deciding whether to pursue it. On seeing the catalogues, however, I realized that this project need not result in frothy and inconclusive discussions or a spurious attempt to marry science and art. Rather, the opportunity to allow the use of *Nature* as a wholly original context for art was intriguing though risky. I was being asked to hand over part of this publication to an artist — in other words, that page of *Nature* (and by association the rest of the

issue) would itself become an original artwork. Other than the right to a blanket rejection if I found the outcome wholly offensive, I would lose any right to editorial intervention.

I agreed to do so. I am no expert in contemporary art, but I was convinced that at least one of the candidates (hitherto unknown to me) was a visually stimulating artist exploring in an original way the associations we make with images and objects — a judgement which, I was delighted to discover, was in line with international recognition and resulted in her being shortlisted for this year's (always controversial) Turner prize. So it is that Cornelia Parker's images, the first set in a weekly series of five, intentionally anonymous and unobtrusive in their presentation, can be found somewhere in this week's News and Views section.

Beneath the surface

Parker is content, for now at least, to let others make what they will of her response to this journal, rather than proclaim her intentions or offer explanations. Indeed, the impact of good art can never be fully explained. On the other hand, objectivity in art (and subjectivity in science) are legitimate and stimulating targets of exploration. Analysts' persistent failure to agree on a harmonic explanation of the famous chord at the start of Wagner's *Tristan and Isolde* may reflect its extraordinarily tantalizing emotive character — one that pervades the entire opera. But we should resist the idea that analysing works of art, and penetrating their objective structure, undermines their capacity to stir the emotions, just as the physicist Richard Feynman refuted the idea that discovering physical laws underlying nature diminishes our sense of wonder at it.

And knowledge of science can help in the appreciation of art, especially where artists themselves have used a scientifically based awareness of the structures and mechanisms of the natural world in generating their more or less subjective creations. It is in that spirit that perceptions shared by scientists and artists will be explored in a weekly series of articles and illustrations discussing art from the Renaissance to the contemporary. That series, by Martin Kemp, professor of art history at the University of Oxford (although originally trained as a scientist), will follow immediately after Parker's.

This, *Nature's* Opinion page, is usually anonymous because its staff believe in the concept of this publication having its own point of view, with final responsibility resting with the Editor. But editorial control has been exerted here to relinquish a small part of the publication to artistic expression. That warrants a more personal acknowledgement of responsibility. It also calls for an acknowledgement that these series are occidental in focus, whereas *Nature's* readership is also drawn from other regions in which discussions of art and science might take a very different direction. So I conclude with an invitation to readers of any origin to send in comments that might usefully broaden joint considerations of art and science, and appreciation of one stimulated by the other.

Philip Campbell



Chip-making equipment industry warns of threat from Intel deal

[WASHINGTON] A \$250-million research collaboration between Intel and the Department of Energy (DOE) to develop new chip-making technology is coming under fire from US suppliers of semiconductor-making equipment, who fear that it will hand over American technology to their Japanese competitors.

Federico Peña, the energy secretary, who announced the agreement last Thursday, will meet the suppliers in Washington this week and, separately, with staff in the Congress. Many of the staff share the equipment suppliers' concerns about the deal, which is the largest Cooperative Research and Development Agreement (CRADA) ever struck by Department of Energy laboratories.

Under the agreement, a consortium led by Intel will pump \$120 million in cash into three national laboratories over three years, and provide a further \$130-million worth of support in kind. The money will be spent advancing extreme ultraviolet (EUV) lithography as a technique for etching microcircuits in silicon.

EUV uses radiation of 13.5 nm wavelength, compared with the visible light wavelengths of 400 nm or more used in existing photolithography machines, potentially etching far more intricate circuits on to a chip.

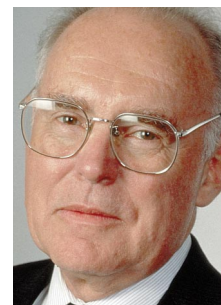
"We very much want [the US equipment suppliers] to be able to participate fully in this venture," says Peña. "We're inviting them to Washington to talk. We're trying to be as sensitive to them as possible — and to get the job done."

But Arthur Zafiropoulos, president of Ultratech of San Jose, California, one of the few remaining US makers of lithography equipment, says that the deal will result in Intel handing over an American technology to the Japanese suppliers — Nikon and Canon — that control almost four-fifths of the lucrative worldwide industry. ASML of Holland holds 15 per cent, and small American suppliers just 7 per cent.

"If this proposal is approved, it will have a devastating impact on the fragile lithography equipment industry in the United States," he says. Ultratech and other US equipment suppliers have previously helped to support the EUV work at the laboratories, he points out. "Intel's proposal will subvert this attempt and destroy the US lithography industry," he says.

The laboratories' expertise in EUV arose in part from the Strategic Defense Initiative, which developed precision optics to manipulate ultraviolet radiation, and partly from nuclear weapons research. In all, 90 per cent of the deal would be shared by two weapons laboratories, Lawrence Livermore National Laboratory in California and the Sandia National Laboratory in New Mexico. The rest goes to the Ernest Lawrence Berkeley National Laboratory in California, where the Advanced Light Source would aid the project.

Intel has sought partners in the microchip industry for the consortium, called EUV Limited Liability Company, which will run the collaboration. But the two other major participants in lithography research,



Intel's Moore: backing ultraviolet lithography.

IBM and Lucent Technologies, have not joined. Advanced Micro Devices (AMD) and Motorola, the two partners who have done so, are said by one source to be investing less than \$1 million each.

Department of Energy laboratories are allowed to do proprietary research for individual companies as well as for consortia, but Congressional staff are nonetheless concerned about several aspects of the EUV CRADA. These include whether Intel will pay the full overhead costs of scientists engaged on the project; the national security implications of releasing the technology, possible future foreign involvement in the consortium, and the intellectual property rights of companies originally involved in the technology.

"It is more than just the concern of a few companies," one staff member says. "We want to know what procedures the DOE has in place to vet an agreement of this size."

As Zafiropoulos points out, the competitiveness implications of the deal are easily large enough to attract Congressional attention. "This is a significant trade issue. We are talking about a \$25-billion market" in the early years of the next century, he says.

IBM researchers continue to back conventional x-ray lithography (with a wavelength of 1 nm) as the successor to light-based techniques, while Lucent is pursuing techniques that use electron beams. "The winner certainly isn't obvious at this time, but we think that EUV is a very strong alternative," says Gordon Moore, the founder of Intel and coiner of "Moore's law", which states that the density of transistors on microchips should double every two years.

The obstacles facing the deal are significant, say congressional staff. But the booming US economy means that continuing Japanese dominance in one key industry — semiconductor manufacturing equipment — is a less potent threat than it would have seemed a decade ago.

The DOE will argue that the deal, which effectively contracts out to the private sector research work that was losing support from the government, closely reflects what Congress has said it wants the laboratories to do.

Colin Macilwain

Brussels slims down advisory body

[MUNICH] The European Commission in Brussels last week approved its restructured science advisory group, the European Science and Technology Assembly (ESTA). The new assembly, whose numbers have been cut from 100 to around 60, with a higher proportion of academic representation than before, will take up its responsibilities next month when the term of the current assembly expires.

ESTA was set up as an advisory group to the commission three years ago, with members primarily from academic institutions and industry. Its reports have included the role of basic research in Europe and the future of nuclear fusion research (see *Nature* 387, 328; 1997).

But the assembly also had teething problems. Lack of administrative support from the commission led last year to the

threatened resignation of its chairman, Jan Borgman, former head of the Dutch National Research Council (see *Nature* 381, 180; 1996). A permanent secretariat has now been established within the commission.

As the number of representatives was found unwieldy, the assembly has been trimmed by a third, and ESTA's 'bureau', the executive group which has 20 members, will be trimmed similarly.

The new assembly will meet for the first time at the beginning of October, when it will elect a chairman who will be ultimately responsible for choosing topics for its future reports. The chairman will also have to consider how to take forward ESTA's most recent report, which sought to identify strengths and weaknesses in Europe's basic research, but was criticized for lack of internal consistency.

Alison Abbott

India opens up its nuclear reactors to foreign inspection

[NEW DELHI] India is to throw open two of its nuclear reactors for inspection by international experts. It has also invited International Atomic Energy Agency (IAEA) specialists to review the design of its new 500-MW reactors — the first invitation in its 40-year nuclear history.

Hans Blix, director general of the IAEA, last week applauded the Indian proposal for international peer review. He said: "It may be an effective method of discovering shortcomings." Blix was in Delhi to inaugurate a world conference on the role of nuclear energy for sustainable development to mark the fiftieth year of India's independence and the fortieth anniversary of the IAEA.

The moves are seen as a response to criticism (see *Nature* 381, 723; 1996) that the country operates vintage reactors of unsafe design. India also wants to dispel the notion it is living in nuclear isolation having refused to sign the Comprehensive Test Ban Treaty.

"Opening up our power plants or subjecting their design to peer review has nothing to do with the IAEA convention on safety of nuclear plants, which India has signed but not ratified," says Yelleswarapu Sivaram Prasad, managing director of the state-owned Nuclear Power Corporation (NPC), which designs, builds and operates the country's nuclear power stations.

"What we are doing is voluntary and is a confidence-building measure," says Prasad, "We want the world nuclear community to find out for themselves that our designs are good and safe."

India has eight 220-MW pressurized heavy water reactors (PHWR) and two US-built 160-MW light-water reactors in operation. Prasad says the Kakrapar nuclear power station in Gujarat, consisting of two PHWR units, will be opened up in November 1997 for visits by members of the Tokyo-based World Association of Nuclear Operators.

"They are free to inspect the control and safety features installed in the power station and watch all the activities going on," Prasad says. "We are not afraid of criticism."

The 220-MW units are based on a standardized Canadian design, but the IAEA team will also review the 500-MW scaled-up version, the first reactor fully designed by NPC. "Should any changes be suggested, we will incorporate them," Prasad says. The first two reactors in the 500-MW series are to be built at Tarapur near Bombay (now Mumbai) with construction to start next year.

Officials hope the open policy may also help India boost its nuclear power production through private-sector participation, both Indian and foreign.

K. S. Jayaraman

Uncertain aims delay start of Internet research effort

[WASHINGTON] The Next Generation Internet (NGI), a \$100-million-a-year research initiative intended to ensure the future leadership of the United States in Internet technology, is getting off to a slow start because of lingering doubts in Congress about its purpose and the geographical distribution of funds to support it.

The programme was announced by President Bill Clinton during his successful re-election campaign last October. It is meant to put those US research agencies that laid the groundwork for what is now the Internet at the forefront of efforts to develop future universal networks of far higher capability.

Almost everyone in the Congress shares that objective. But after initial confusion about the nature of the programme, various congressional committees have been reluctant to provide money for it. As a result, NGI may stagger into 1998 with only half the \$105 million requested by Clinton for its first year.

NGI has been designed by the administration to carry out the research needed to develop the Internet of the future. It is distinct from Internet 2, a project by a hundred research universities in the United States to establish a fast network for their own exclusive use to bypass the clogged arteries of the existing Internet.

The popularity of the Internet in the United States — bolstered by unlimited domestic access via free local telephone lines — has made it slower there than in many other countries.

The administration has divided NGI into three parts. The first, to be led by the Defence Advanced Research Projects Agency (DARPA), is intended to develop technologies and protocols for a much faster Internet.

The second, according to the project plan, will "weave the next generation network fabric" by developing a test bed of 100 linked sites for the new network. The administration wanted the Department of Energy (DOE) to lead this, but is now coming round to Congress's view that the National Science Foundation (NSF) should lead it.

The third element, which will involve several government agencies, would develop

US funding proposals for the Next Generation Internet research initiative

Agency	Clinton request	House approp.	Senate approp.
DARPA	\$40m	\$40m	\$10m
DOE	\$35m	0	0
NSF	\$10m	\$23m	\$10m
NASA	\$10m	\$10m	\$10m
NIST	\$5m	\$5m	\$5m
NIH	\$5m	\$5m	\$5m
Total	\$105m	\$83m	\$40m

"revolutionary applications" for the new network. But by the time details of this plan were finalized in July, congressional appropriations panels had shunned the initiative.

Congressional staff of both parties lay the blame for this at the door of the Office of Science and Technology Policy (OSTP), the White House science office responsible for coordinating the initiative across the federal government. They say that OSTP was too slow to formulate details of the proposal after Clinton announced it.

"The train left the station before we had any realistic idea of where it was going," says James Sensenbrenner (Republican, Wisconsin), who chairs the Science committee in the House of Representatives. The committee failed to approve any money for the programme when it recommended funding levels for research programmes under its jurisdiction back in March.

Complaints then surfaced from powerful senators for rural states, notably Ted Stevens (Republican, Alaska), who chairs the appropriations committee, and Conrad Burns (Republican, Montana), who feared that the initiative would discriminate against remote locations. The NSF has since offered \$200,000 to help some rural states gain access to high-speed links. But at a hearing of the House Science committee last week, Joe Thompson of Mississippi State University said that this would cover "less than 10 per cent" of the extra costs these states face.

When appropriations subcommittees considered NGI funding in June and July, the patchy outcome reflected both geographical and organizational concerns (see table). This month, the House and the Senate will compromise on the funding levels for this and other programmes at each agency. Sensenbrenner says NGI is likely to end up with "50 to 60 per cent" of the \$105 million the administration wanted for it in 1998.

Since the plan was published in July, the NGI initiative has won the support of many key players in the Congress, and is likely to be fully funded in 1999. As the purpose of the initiative becomes clearer, and the administration moves to accommodate the concern of states that fear being left behind, support should solidify.

But the delay in starting the initiative could dent its relevance to a field of technology which moves more quickly than the funding system can handle. As Henry Kelly, acting associate director of OSTP, told the hearing: "The danger here is that we'll have a slow start, which is an unfortunate situation when you're dealing with a technology that moves like lightning."

Colin Macilwain

Science to stay on top at French museum

[PARIS] France is set to abandon controversial plans to turn the world's largest anthropology museum — the Musée de l'Homme in Paris — into a museum of fine arts. The most likely result, a rebuff for the neo-Gaullist French president Jacques Chirac, who staunchly supported the idea, will restore the museum's scientific role while retaining an artistic dimension to exhibitions.

The Musée de l'Homme has fallen into decay over the past few decades, and Chirac's decision last year to give it a multimillion-dollar facelift was broadly welcomed. But the renovation plans have been plagued by controversy, as scientists and art specialists have fought bitterly over competing visions for the museum (see *Nature* 383, 109 & 203; 1996).

Until now, art specialists have had the upper hand. Jacques Kerchache, an art collector and a former dealer in African art, was influential in persuading Chirac to support plans to convert the Musée into a museum of primitive art. Kerchache is a close friend of Chirac, who is himself a connoisseur of early art.

Kerchache's ideas were endorsed last year by a commission appointed by Chirac which recommended the creation of a 'Museum of Mankind, Art and Civilization'; this would have merged the millions of objects held in the collections of the Musée de l'Homme and the Museum of African and Oceanic Arts.

But the commission's recommendations were fiercely criticized by the scientific community. Many felt that the transfer of responsibility for the ethnology collections from scientists to art curators would result in objects being presented as 'art', devoid of their cultural context.

Under the plans, responsibility for the museum, formerly exclusively that of the science ministry, would be shared with the ministry of culture. Many predicted that this would result in a watering-down of the museum's research activities.

Defending a broader vision

The commission was dominated by art specialists, including Kerchache — only four of its sixteen members were scientists — and it was widely perceived as a 'rubber-stamp' affair. In particular, the commission rejected a FF400-million (US\$67-million) renovation programme proposed by Henry de Lumley, director of the museum's parent body, the Natural History Museum, which would have expanded the museum's research capacity, and created science-based exhibitions showing the sweep of mankind's social, cultural and biological evolution. De Lumley's defence of a broader vision for the museum was strongly backed by museum scientists worldwide, as well as personalities including Nelson Mandela.

The election of a socialist government in



Art versus science: Proponents of both have been in conflict over the Musée de l'Homme (right) in Paris.

June has radically changed the landscape, however, admits Christine Albanel, Chirac's adviser for education and culture. It means that a decision on the museum's future must now be agreed by both Chirac and Lionel Jospin, the socialist prime minister. According to Bertrand Mabillet, Jospin's adviser for research, technology and space, the government considers that science has been neglected in the present plans for the museum; it will submit its proposals to Chirac in the coming weeks.

Mabillet says the government is keen to find a compromise acceptable both to Chirac and to scientists and art specialists. As a step in this direction, it intends to propose that scientists be given a greater say in the association set up by the previous conservative government to work out detailed plans for the new museum.

The association has been particularly criticized because of the presence on its scientific board of Kerchache and two other art dealers, Georges Baselitz and Jean-Paul Barbier, who recently sold specimens of primitive art to the state for FF40 million. De Lumley argues that there is a conflict of interest in having art dealers involved in policy making, particularly in matters concerning new acquisitions.

But de Lumley's arguments are rejected by Germain Viatte, the board member in charge of the project. A former director of the National Museum of Modern Art at the Centre Pompidou, Viatte defends the inclusion of art dealers on the scientific board. "I put them there because they are collectors who know how to create a museum and exhibitions," he says, adding that scientists are naive in thinking that museums can survive without getting involved with the commercial art world.

Mabillet says the government believes that the membership of the association needs to be "re-equilibrated" by bringing in more scientists, otherwise its recommendations risk being "too *beaux arts*". As a first step, the government intends to propose that a leading scientist be appointed alongside Viatte in an

attempt to achieve a "consensus" between the demands of art and science.

This proposal is opposed by Viatte, but supported by the Elysée, the president's office, according to Albanel. A separate scientific advisory group may also be established to design a scientific strategy for the museum, says Mabillet, adding that "research and teaching must exist alongside art".

Compromise within reach?

Viatte argues that the conflict between a scientific and an artistic approach is an "old debate with no sense", and that there is no reason why an artistic vision of the museum cannot coexist with a scientific one. But he admits that science was ignored in the original proposals for the new museum.

According to Viatte, an artistic aspect is "fundamental" to the museum, and is "completely absent" in the existing Musée de l'Homme. He argues that de Lumley's proposals neglected this aspect, and amounted to little more than a "school textbook". For his part, de Lumley says he is no longer opposed to an artistic dimension, provided the museum's scientific mission is retained.

The final shape of the museum will depend heavily on a decision on whether to move the Musée de la Marine, which shares the same buildings as the Musée de l'Homme, to another location. The original proposal, to move the museum to a site outside Paris, has now been abandoned as too costly, with estimates running as high as FF1.5 billion.

The government has set up a commission to decide on the site, and its recommendation is due next month. If the museum were to move, it would probably be built from scratch on a vacant site on the banks of the Seine. De Lumley believes the Musée de l'Homme could be renovated without moving the Musée de la Marine, and that this could reduce the costs to FF300–400 million. But Viatte disagrees, arguing that only by moving the museum can sufficient space be freed to allow an ambitious project.

Declan Butler

NASA looks to overhaul grants process

[WASHINGTON] NASA, the US National Aeronautics and Space Administration, has embarked on a sweeping review of its research grants programme, aiming to make it more efficient and responsive to its own needs, and to allay the academic concern about prompt grant payments.

The review, which began last month and will run through the winter, will look at the space agency's entire grants process, from the inception of an idea for a research solicitation to peer review, grant awards, and final work performance and payment. It is being led by Mary Kicza, associate director for space science programmes at the Goddard Space Flight Center in Maryland.

Kicza, who formerly ran the agency's Discovery programme for low-cost planetary missions, says NASA knew it had bottlenecks and other problems in its grants administration. But recent changes have not helped. At the same time as NASA headquarters has been cutting staff, responsibility for administering research grants shifted from there to Goddard, which has had to hire contract officers and learn new procedures.

Meanwhile, pressure to reform is coming from inside and outside the agency. NASA's administrator, Daniel Goldin, for example, has spoken of changing the peer review system to encourage young scientists and unconventional ideas. The White House Office of Science and Technology Policy is eager to improve relations between the government and the universities, and is paying close attention, as is Congress, to complaints from the academic community about NASA's grants process.

One frequent complaint involves delays in payment. Kenneth Baldwin, professor of physiology and biophysics at the University of California at Irvine, and a member of NASA's advisory committee for life and microgravity sciences, said last week that whereas the National Institutes of Health (NIH) generally take three weeks or so to give an initial report on a grant application's scientific merit, NASA can take as long as three months. Other NASA-sponsored scientists report that cheques arrive late, sometimes well after the funding year has begun.

NASA's recent efforts to cut its 'uncosted carryover' budget — approved funds left unspent before the end of a fiscal year, which roll over into the next year — have also imposed hardships on universities, say Kicza and others. In effect, if NASA chooses not to bridge the gap from one funding cycle to the next, universities have to do so.

Life and microgravity scientists, on whom NASA spends about \$114 million a year for 740 grants, have been hit particularly hard by this problem. Because agency programme managers face a funding shortfall in these areas next year, they have been particularly aggressive about reducing their uncosted carryover.

Kicza says her team is eager to hear complaints from the scientific community, and has sent out thousands of letters to individuals and academic institutions to solicit comments. Her office also has a Web site (<http://booster.nasa.gov:443/grant>) for researchers to fill out a questionnaire online.

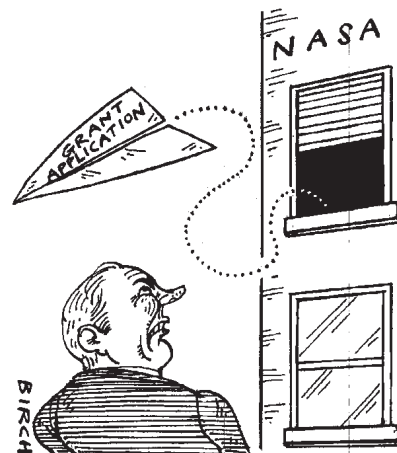
At present, she says, the study team is simply taking stock of how NASA's different

offices, which fund research in disciplines ranging from astronomy to biology to fundamental physics, process their grants. By early next year, however, they hope to have solutions.

One obvious area to look at, she says, is automation of grant-processing. The National Science Foundation has developed a system whereby grant applicants can send proposals electronically, and the NIH should have a system in place by 1999. An Interagency Electronic Grants Committee of the Government Services Agency is looking at standardizing techniques for all government granting agencies.

Kicza says her group will consult frequently with other agencies that process grants to see what NASA can learn. Her team's report will be presented to Goldin around March, in time to influence the 1999 budget request.

Tony Reichhardt



Lack of cash for instruments 'threatens ESA science missions'

[MUNICH] Future scientific missions of the European Space Agency (ESA) are under threat because member states are not making sufficient money available for the development of scientific instruments.

This was the message from a 'brainstorming' meeting in Oxford, England, last week of ESA's science advisory and science programme committees, and its working groups on astronomy, Solar System science and fundamental physics.

Under ESA rules, the agency pays for satellites and launches but member states must provide instruments for missions in which their scientists participate. The meeting was arranged to discuss responses to financial crises faced by both sides.

ESA has reoriented its long-term space science plan, Horizons 2000. The new plan, presented to the meeting by the agency's executive, abandons the concept of ECU300-million (US\$332-million) medium-sized

missions in favour of smaller and cheaper 'flexible' missions, including technology testing missions, that can be launched more frequently (see *Nature* 385, 380; 1997).

ESA's executive said it also hopes to make a one-off saving of more than ECU300 million by placing instruments for the next two missions in the planning stage — the cornerstone FIRST (Far-Infrared Telescope) and the Planck surveyor, a medium-sized mission to study the cosmic-microwave background — on the same satellite (see *Nature* 387, 639; 1997). The merger is strongly opposed by Planck scientists.

Ironically, says Paul Murdin of the UK Particle Physics and Astronomy Research Council, ESA's attempts to make its science budget stretch further will make it even harder for member states to provide instruments. Delegates said it would be difficult to find cash for instruments for early flexible missions such as Mars express,

which ESA hopes will be launched in 2003. In addition, the proposed FIRST-Planck merger, with a planned launch in 2005 or 2006, would bring forward FIRST's launch by two years, putting further pressure on budgets. A firm decision on the merger will be made next June. Meanwhile, a call for proposals for FIRST and Planck instruments will go out in the next few weeks.

No definitive solution to the cash problems was proposed, but "many options for alleviating costs to member states were discussed", says Roger Bonnet, ESA's science director. These included simplifying instruments, and standardizing interfaces between instruments and satellites so costs could be transferred to ESA.

There was some lightening of the gloom with the announcement that NASA would provide a 3.5-metre carbon-fibre telescope for FIRST as a demonstrator in return for 10 per cent of viewing time.

Alison Abbott

Ozone treaty 'must tackle CFC smuggling'

[MONTREAL] Representatives of more than a hundred signatory countries to the Montreal Protocol on Substances that Deplete the Ozone Layer gathered in Montreal this week to celebrate the signing of the protocol exactly 10 years ago, and to consider ways of strengthening it.

As in previous annual meetings, there are likely to be measures to accelerate the phase-out of ozone-depleting substances, and this week measures to discourage illegal trade in chlorofluorocarbons (CFCs) will be high on the agenda. Exemptions for what are considered essential uses for certain chemicals will also again be considered.

Virtually all production and import of CFCs has been banned in the United States and other signatory countries, but sales are not illegal. As a result, smuggling of the chemicals into the United States from developing countries, where production is still allowed, has become big business, with huge profits to be made.

From 1994 to 1996, an estimated 10,000 tonnes were smuggled into Florida, where its street value is thought to be not much lower than that of cocaine. On the Mexican border, a CFC-containing coolant was reported last November to be second only to marijuana in the value of contraband seizures.

Black-market demand is high because owners of older vehicles are unwilling to pay the much higher price for substitutes that can be used in air-conditioning. A 12-ounce container of CFC-containing coolant sells for US\$2 in Mexico and \$23 in California.

Thomas Watts-Fitzgerald, a US attorney involved in a programme to stop CFC imports, says that while converting cocaine to crack and selling it brings a fourfold profit, the sale of a canister of Freon, a DuPont brand-name coolant, can bring a 13-fold profit.

Signatories of the Montreal Protocol are now being asked by the Environmental Investigation Agency, a conservation trust based in London and Washington, to adopt a licensing system that could track all CFC imports and a ban on unlicensed imports and exports of controlled substances.

The Montreal Protocol was signed by 24 countries on 16 September 1987, and has 163 signatories. According to its Web page (www.ec.gc.ca/MontrealProtocol), the protocol is "the first international mechanism designed to address an arising global environmental problem".

The theory that CFCs — then used in refrigeration, aerosol cans and some industrial processes — were implicated in destruction of the ozone layer was first proposed by two University of California scientists, Sherwood Rowland and Mario Molina, in the 1970s.

In 1977, the United Nations Environment Programme (UNEP) established a



Leakage? Illegal trade in legally produced CFCs, here being bottled in India, offers huge profits.

Coordinating Committee on the Ozone Layer and adopted a World Plan of Action. In the late 1970s and early 1980s, the United States, Canada and other countries banned CFCs as aerosol propellants for non-essential uses, reducing CFC consumption in Canada's case by more than 50 per cent.

In 1985, Canada became the first signatory of the Vienna Convention on the Protection of the Ozone Layer, which two-and-a-half years later became the Montreal Protocol. The protocol came into effect on 1 January 1989. Signatories agreed to freeze consumption of CFCs and halons at 1986 levels, and then to reduce consumption by 50 per cent within 10 years. Developing countries were given a grace period of 10 years.

Since then, adjustments have been made to the scope and stringency of the controls, and further substances, such as carbon tetra-

chloride, methyl chloroform, HCFCs, HBCFs, and methyl bromide, have been added to the regulated list. One result has been a dramatic lowering of the use of CFCs and halons: by 90 per cent in OECD countries between 1986 and 1994.

Accelerated phase-out of methyl bromide for industrial countries — it is used as a pesticide in agrifood businesses worldwide — will be under discussion at the Montreal meeting. Molecule-for-molecule, methyl bromide is considered at least 50 times more destructive to the ozone layer than chlorine from CFCs.

Also due to be discussed is a phase-out schedule for developing countries and exemptions for the chemical where there are no technically or economically feasible alternatives. Delegates will debate whether and how to disallow trade in the substance with countries that are not parties to the protocol.

A conference of environmental non-governmental organizations (NGOs) is being held in conjunction with the main meeting, with speakers and panels on an ozone recovery plan and assessments of the performance of the protocol. There is also a Virtual Environmental NGO Conference at www.intranet.ca/~foe/.

The Canadian federal government and UNEP will present awards at an anniversary ceremony. Canada's leading role in the adoption of the protocol is just one factor in this country's interest in ozone depletion. A recent scientific survey from its scientists points out that ozone thinning continues over Canada. "In the high Arctic, severe and unexpected losses of up to 45 per cent have occurred in the springtime," the report says. "UV [ultraviolet] levels are expected to remain higher than normal for 30–40 years."

David Spurgeon

'Surprising success' of the Montreal Protocol

[MONTREAL] One of the first scientists to have pointed out in the mid-1970s the potential threat to the atmosphere of CFCs said last week that he had been both surprised and impressed by the success of the Montreal Protocol in restricting their emissions.

"It was frustrating for many years, but it really paid off with the protocol, which was a marvellous example of what the international community can do working together," said Mario Molina of the Massachusetts Institute of Technology, who with Sherwood Rowland was co-author of the research paper

published in 1974 that first drew attention to the problem.

"We can see from the atmospheric measurements that it is already working," Molina said in Montreal last week, where he was at a meeting to mark the tenth anniversary of the protocol. "It has imperfections, but the fact is that we have already found that the concentration of chlorine getting into the stratosphere has levelled off and is starting to decrease."

Molina says that, in hindsight, it is always possible to find better ways of doing things. "But I am still amazed that the protocol

worked; for many years I thought that it was hopeless — how were we going to get industry and the different governments in developing countries to agree to stop this? What I think helped was on the one hand the very clear scientific evidence and on the other the very clever diplomatic negotiations."

In terms of the science, Molina says that it was a "very small community" that dealt with this issue. "One very important component is capacity building — we need to have more scientists in the developing world working on these issues."

D.S.

Australia eyes role in Gemini telescopes

[SYDNEY] Australia may, after all, be involved in the Gemini project to build twin telescopes in Hawaii and Chile, even though its chance of being a replacement for Chile ended when the Chilean government finally agreed to pay its fee just before the 1 September deadline set by the other Gemini countries.

At a meeting of the Anglo-Australian Observatory (AAO) board in Durham, England, last week, Australian representatives gained backing for their efforts from British members. Ian Corbett, director of science for

the UK Particle Physics and Astronomy Research Council, and a member of both the AAO and Gemini boards, says he is "very hopeful that a solution can be found".

"We all recognize the benefits to Gemini if Australia were to join and contribute its expertise and additional funds," says Corbett. "We are therefore now working together to see if we can establish a basis on which Australia could be invited to join the collaboration."

Roger Bell, the AAO secretary, said after his return to Sydney this week that negotiations had begun on how Australia might join

as an additional partner. "Australia has come up with the money," he says, adding that while AAO itself cannot be a negotiating or funding party, a meeting between Gemini and Australian representatives could be arranged "soon".

Australia had an earlier opportunity for sharing in a large international telescope when invited by the European Southern Observatory (ESO) in 1994 to become its first non-European member in the Very Large Telescope (VLT), a set of four linked 8-metre instruments under construction in Chile.

The bid was foiled last year, however, when, shortly after the federal election, the new science minister, Peter McGauran, refused to commit the necessary funds of A\$25 million (US\$18.5 million) over six years while the conservative Coalition government was focusing on cutting expenditure (see *Nature* 381, 100; 1996).

Last May a new opportunity emerged when the Gemini consortium building the two 8-metre telescopes for US\$184 million approached Australia as a potential replacement for Chile. The Chilean government had blocked its cash share of Gemini along with separate contributions due for ESO and the US telescopes in the Andes.

The Gemini consortium — which includes the United States (50 per cent), the United Kingdom (25 per cent), Canada (15 per cent) and Brazil and Argentina (2.5 per cent each) — set a deadline of 1 September for Chile to pay its 5 per cent. The fee for Australia to join Gemini in place of Chile — A\$12 million over four years, with A\$4.7 million paid up-front — was less than its ESO membership fee.

While McGauran appeared to be supportive this time, despite a climate of continuing cuts in public expenditure, funds in his portfolio were consumed by research agencies and centres. Astronomers shifted their lobbying to the education minister, Amanda Vanstone, then found a champion in Max Brennan, a plasma physicist who then chaired the Australian Research Council (ARC).

Australia negotiated through a national consortium led by Jeremy Mould, director of the Mt Stromlo and Siding Spring Observatories of the Australian National University; the university runs three telescopes in New South Wales on the same site as the 3.9-metre Anglo-Australian Telescope operated by AAO.

While government officials expressed confidence, just before the Chilean deadline, that there would be an imminent "positive announcement" on funding, a letter of commitment had apparently already been sent from the ARC to the Gemini board in early August. Vicki Sara, newly appointed successor to Brennan (see panel), confirms the ARC's continuing support. **Peter Pockley**

New research chief faces multiple challenges

[SYDNEY] Vicki Sara, newly appointed as chair of the Australian Research Council (ARC) and the first woman in the post, says her first challenge is to ensure Australia's international competitiveness in research at a time of "extreme pressure on funding".

The council provides competitive, peer-reviewed grants for research in universities. Appointed for three years full-time, Sara succeeds Max Brennan, a physicist, who retired in August.

Brennan's departure had been signalled 12 months earlier, and the long delay by the Education Minister, Amanda Vanstone, in announcing a successor was publicly attacked by the country's four learned academies and the Australian vice-chancellors' committee, which were concerned that it presaged major changes to the ARC.

Sara's appointment has now been widely welcomed by these bodies. But the delay will cause a four-month hiatus, as she cannot leave her current job as dean of science at the Queensland University of Technology (QUT) until the beginning of next year.

Sara, who is 51, gained a high reputation for her research in the endocrinology of fetal brain development during her 17 years at the Karolinska



Sara: will encounter political tensions in a climate of cuts.

Institute in Stockholm, Sweden. After returning to Australia in 1993, she became the first director of the Cooperative Research Centre for Diagnostic Technologies, based at QUT.

On moving to Canberra, she will enter a tense political atmosphere characterized by continuing conflict between Vanstone and university researchers over the effects of cuts. Anxiety about the government's responses to recent reviews was heightened by Vanstone's statement, when announcing Sara's appointment, that the ARC will soon be making "critical funding and programme decisions".

These decisions may have to be made before Sara can develop the ARC's first strategic plan, which she says will take six months.

Next May, the Coalition government delivers its third and last annual budget before the next federal election. The 1997 budget foreshadowed substantial cuts to research funding — ARC's budget will drop from

A\$429 million (US\$309 million) in 1997–98 to A\$381 million in 2000–01 (see *Nature* 387, 222 & 328; 1997).

Sara acknowledges that this situation poses a major challenge. "We must never give up the aim of improving our funding," she says, expressing particular concern at the effect on careers and international competitiveness of the fact that the ARC must "provide for an infrastructure on 27 cents in every dollar of research funding, whereas our aim should be 40 cents".

She believes the success rate in grant applications (21 per cent) is "too low", citing a "minimum rate of around 30 per cent in Europe".

Encouraging more cooperation between institutions and more sharing of major international facilities, such as Australian astronomers' attempt to join the Gemini telescopes project (see above), which she strongly supports, are Sara's first suggestions for alleviating the effects of cuts.

She is the first chair of the ARC to come to the post declaring a need "to boost the profile of research in the public and politics" through the ARC. She is an active promoter of the public understanding of science, having been involved in the 'QUT Science Train', an educational show for young people that is at present touring Queensland. **P.P.**

Image of Newton reopens historic divide

[LONDON] The unveiling last week of a 12-foot-high, bronze statue of the seventeenth century physicist Sir Isaac Newton (right), outside the British Library's new building in central London, has renewed a debate over the significance of the painting by the eighteenth century poet and artist William Blake (below right), on which the design of the statue is based.

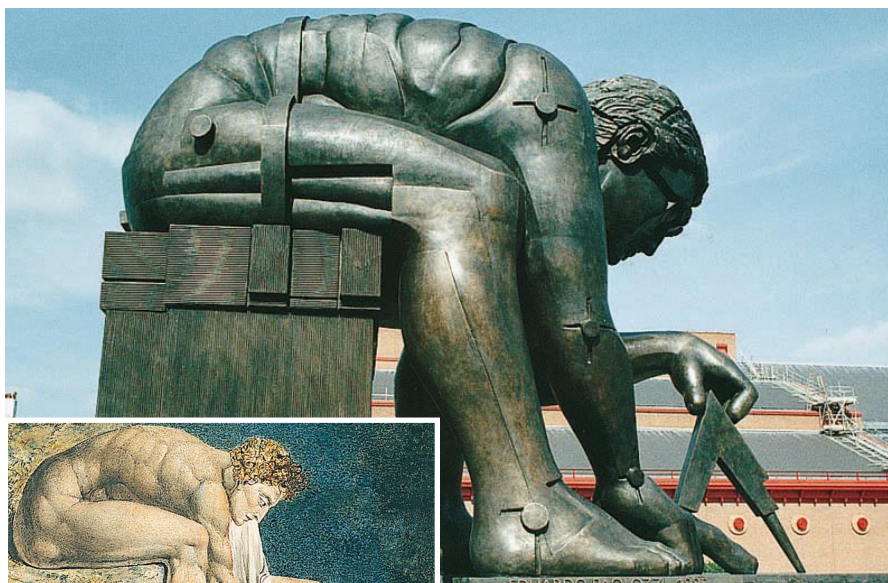
Sculptor Sir Eduardo Paolozzi says his statue is intended to show how art and science are interconnected. But at least one historian says the choice of subject is inappropriate, as Blake was highly critical of the English physicist.

Both the £175,000 (US\$280,000) statue, as well as Blake's original painting of Newton, show the physicist bending forward to plot or measure the Universe.

Simon Schaffer, reader in the history and philosophy of science at the University of Cambridge, says that this image was not designed to depict Newton in a flattering light. "Blake is associating Newton with an image of God as a tyrannical, and powerful measurer of the world. He is not celebrating Newton at all."

But Paolozzi, an honorary professor at the Royal College of Art in London, says: "If you see the original it shows that Blake made Newton into a God. He was extremely respectful of [Newton's] abilities." Paolozzi says that the imagery in the painting is at worst "a slight dig" at Newton.

Schaffer points out that Blake was among the first poets who attempted to understand



the implications of Newton's natural philosophy at a time when most contemporaries were happy merely to celebrate the triumph of reason. He says Blake disapproved of the view that science and reason were sufficient in themselves to explain the workings of the Universe, or those of life on Earth.

Blake, says Schaffer, was also critical of libraries, which he felt encouraged knowledge to be accumulated for display,

instead of being put to practical use.

But one leading art historian says the statue's concept is neither unusual nor inconsistent. It follows an established tradition in art in which the founders of mathematical sciences — Archimedes, Ptolemy, and Euclid — are shown measuring the world.

Ken Shirreffs, a spokesman for the British Library, says all those involved in the project are "aware of the ambiguities of Blake's views regarding Newton". But he adds that the statue has a much wider significance. "It is a fusion of art and science," and is an appropriate symbol for the British Library. The library opens in November. **Ehsan Masood**

C. P. Snow's 'two cultures' thesis was just 'declinist whinging'

[LEEDS, ENGLAND] Claims by the late British novelist C. P. Snow that a 'two cultures' gap separates those who have studied the humanities from those who have studied the sciences came under spirited attack last week from a historian of science for misrepresenting the central role science and technology have long played in British life.

"Snow is part of the problem, not the solution," David Edgerton of Imperial College, London, told the annual meeting of the British Association for the Advancement of Science in Leeds. He described Snow as an example of "declinist whinging" about British culture which remains "very popular among scientists and engineers".

Snow trained and worked as a physicist before the Second World War, later becoming a prominent novelist, a civil servant advising on scientific recruitment, and eventually a government minister.

His 'two cultures' thesis, presented in a lecture given in Cambridge in the late 1950s, remains widely quoted. Snow argued that

British culture, in contrast to that of its main economic competitors, had traditionally been 'anti-scientific' and opposed to technical progress. The lecture itself is still in print.

But, Edgerton points out, Snow "has no explanation for the rise of British science", nor for the 50-fold increase in the number of scientists in Britain between 1902 and 1966. "Snow is invoking an England without Charles Babbage, Michael Faraday, James Joule or Lord Kelvin," he says. "Indeed without anything that could possibly explain that creation and rise of new universities with their strong emphasis on science and engineering, or the continued growth of British science-based industry."

Edgerton also criticizes Snow's international comparisons, suggesting that there is also reason to doubt his description of Britain's apparent weakness. "Britain may have been deficient in turning out graduates in science and engineering compared to the USA and the USSR, but not France,

Germany or Japan," he says.

Snow's characterization of the 'two cultures' gap, built largely on the accepted fact that most senior politicians and civil servants came from humanities backgrounds — while many academics from such disciplines were often disparaging about their scientific colleagues — remains the starting point for much discussion about issues such as the future of Britain's education system.

But Edgerton complains that Snow's vision is based almost entirely on attitudes towards academic research physicists, while the views taken to represent British culture as a whole are predominantly those of British novelists.

"The fact that Snow has been taken seriously is a testimony to the importance of science in British culture," said Edgerton. "A modern justification for the [study of] the history of British science is not that Snow was right, but that he was profoundly wrong."

'Science endowment should take risks', UK minister told

[LONDON] The UK government's proposed National Endowment for Science, Technology and the Arts (NESTA) should take risks and not be fearful of public reaction to its choices of beneficiaries. That was the call from representatives of the arts, crafts, engineering and sciences at a meeting held in London last week to mark the close of national consultations about NESTA's role (see *Nature* 386, 749; 1997).

Chris Smith, minister for heritage and arts, emphasized Britain needed to make wealth out of its "creative industries". But some delegates, including Britain's most recent Nobel laureate, Sir Harry Kroto, urged that the endowment be used to stimulate artistic and scientific creativity free of both wealth-creation aims and the bureaucratic burdens of accountability normally associated with government expenditure.

The government was also urged not to make NESTA, funded from the National Lottery, dependent on income from intellectual property rights, which should instead go to innovators. Criticisms were expressed that consultations suffered from a lack of detail. The government will now draft legislation to establish NESTA next spring.

Magnetic scattering device opens at ESRF

[LONDON] A £2-million instrument enabling British scientists to study the atomic and magnetic properties of magnetic materials will be opened tomorrow (19 September) at the European Synchrotron Radiation Facility in Grenoble, France. The X-Ray Magnetic Scattering device (known as XMaS) took four years to build and involved scientists from the Universities of Warwick, Keele and Liverpool. It was funded by the Engineering and Physical Sciences Research Council, which has also awarded £2.3-million to the universities of Liverpool and Warwick for a five-year programme of experiments.

Congress urged to act on medical privacy

[WASHINGTON] The US government last week called on Congress to enact a national law that would make it a criminal offence to improperly obtain or divulge protected medical information (see *Nature* 388, 611; 1997). Donna Shalala, the Secretary of Health and Human Services, sent Congress an 81-page report asking it to impose criminal and civil penalties on doctors, hospitals, researchers, insurance companies and others who fail to adequately safeguard identifiable patient

information or who obtain it improperly.

The proposed law would also allow patients the right to see their records and discover who else has seen them. Researchers conducting blinded trials, however, would be allowed to ask patients to waive this right. Researchers would also retain access to identifiable records without patient consent if local ethical review boards approved.

'Consensus conference' on human genetics

[LEEDS, ENGLAND] The British government is to mount a 'consensus conference' on human genetics, designed to help promote public agreement on key aspects of modern genetics and their social implications. John Battle, the minister for science, energy and industry, said last week that the meeting, which is still at an early planning stage, is to be organized in collaboration with the Human Genetics Advisory Commission, set up last year to monitor such issues.

Speaking during the annual meeting of the British Association for the Advancement of Science in Leeds, Battle said he was keen to stimulate greater involvement by the public in debates about science as part of a broad move to "bring science in from the margins" of modern culture. But he also wanted to ensure the genetics meeting "is a consensus conference and not a conflictual one".

Lazar appointed to lead research cooperation

[PARIS] Philippe Lazar, an epidemiologist who was president of INSERM, the French national biomedical research organization, from 1982 to 1996, has been appointed president of the board of ORSTOM, France's agency for research cooperation with developing countries. Jean Nemo will remain the agency's director-general.

A spokeswoman for the agency welcomed Lazar's appointment as "good news", as it suggests the government does not intend to abolish the agency, a move that was considered in July (see *Nature* 388, 7; 1997). But the spokeswoman said Lazar had made it clear he had no intention of taking up the post if it meant overseeing the "liquidation" of the agency.

German president hosts 'science garden party'

[MUNICH] Roman Herzog, the German president, last week held a heavily publicized 'Fest der Ideen' — festival of ideas — in the garden of his Berlin home, inviting 5,000 inventors and researchers to present their ideas, as a way of encouraging creativity in science and technology. The diverse and often highly photogenic products included a self-steering balloon, a machine that takes the top

off boiled eggs, a pocket-size laser which can be used for medical purposes, and graffiti-resistant paint.

In the past few months, Herzog has been actively campaigning for more risk-taking and imagination in German society, in order to stimulate the long-term growth of the economy. There is too much pessimism about the use of new technologies, he told the gathering in his garden. "Germans have been discussing for too long only risks and dangers," he said. "But missed opportunities are a risk as well."

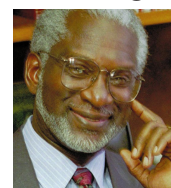
Eaton plans to quit geological agency

[WASHINGTON] Gordon Eaton, the director of the US Geological Survey, announced last week that he would leave the agency on 1 October. The 68-year-old Eaton joined the survey in 1994, and his tenure was marked by reorganization and downsizing, including the absorption of the National Biological Service, which became the USGS Biological Resources Division. Most recently, Eaton handed down a controversial decision to relocate the agency's offices in Menlo Park, California (see *Nature* 389, 3; 1997).

Fallout from that move is not believed to have played a major role in his decision to leave, however. In a farewell letter to employees, Eaton wrote: "From the

beginning it was understood that my tenure would be relatively short, for my 65th birthday was behind me when I took office." Interior Secretary Bruce Babbitt is expected to name an acting director "soon" while searching for a permanent replacement.

Satcher nominated as US Surgeon General



[WASHINGTON] President Bill Clinton last week nominated David Satcher, director of the Centers for Disease Control and Prevention (CDCP), to be Surgeon General, vacant since 1994. Clinton also nominated Satcher, an expert in sickle cell disease, to be assistant secretary for health at the Department of Health and Human Services. This would add a policy-making responsibility to the Surgeon General's role as the country's top public health advocate.

Satcher, who is 56, has headed the CDCP since 1993. Before that, he was president of Meharry Medical College, a historically black medical school in Nashville, Tennessee. Satcher pushed for Clinton's apology this year to victims of the Tuskegee experiment, in which black men suffering from syphilis in Alabama, Satcher's home state, were left untreated (see *Nature* 387, 116; 1997).

Italy needs its an NIH of its own

Sir — Italian scientists must be grateful to *Nature* and to Alison Abbott for providing an international forum to debate an issue that many Italians consider 'an internal affair' to be handled in the usual byzantine way (*Nature* 388, 609–610; 1997). I would like to comment on two points concerning the future of biomedicine in Italy. The question of the future shape and responsibilities of the Higher Institute of Public Health (Istituto Superiore di Sanità, or ISS) is complex because, as Abbott points out, the institute is burdened by a number of regulatory duties. It is not opposed in principle to the transfer of these tasks to new *ad hoc* bodies, similar to the US Food and Drug Administration (FDA) or Environmental Protection Agency (EPA).

Our fear is, however, that the government may create new agencies, leaving all the technical work to be carried out by the ISS because it would be too difficult to find enough new experts and too costly to build and equip the infrastructure necessary to undertake all the controls exercised by an FDA or EPA.

What the government must do is to choose between two models: to create new agencies and provide them with the funds, expertise and infrastructure needed to carry out their duties independently or, alternatively, to restructure regulatory duties within the ISS in such a way that they become as separate as possible from research activities. A hybrid solution would help no one.

Biomedical research in most Western countries, contrary to what happens in Italy, is supervised by a body separate from the National Research Council. In the United States, federal biomedical research is carried out by the National Institutes of Health (NIH), which are directly responsible to the Department of Health and Human Services; health-related research in the United Kingdom and Scandinavia comes under the Medical Research Councils of those countries, and in France under INSERM.

Combining the ISS with Italy's network of 22 national clinical research institutes (Istituti di Ricovero e Cura a Carattere

Scientifico), to act as a 'national institute for biomedical and health research', will not by itself take anything away from the national research council (CNR) as some fear. Funds for the new entity would continue to come from the Ministry of Health, and the CNR, which is funded by the Ministry of Research, would continue to have its own network (the Committee for Biology and Medicine and CNR Medical Institutes).

The relationship between the two networks needs to be resolved but, at least in principle, an 'Italian NIH' should carry out research directly relevant to the activities of the national health system, whereas the CNR could concentrate on more basic aspects.

What is needed above all is to avoid waste, duplication and the splitting of funds into too many channels that benefit only 'clients', not researchers.

Giuseppe Benagiano

(Director General)

Istituto Superiore di Sanità,
Roma, Italy

Cheques and balances

Sir — Sir Hermann Bondi rightly remarks that there must be an optimum balance or ratio (albeit subject-dependent) between the expenditure on scientists and that on their equipment, and points out that the current ratio is too high because of the widespread obsolescence of equipment (*Nature* 388, 709; 1997). His solution, however, is that "the claimant pool [of scientists] should be set firmly on a downward course".

The alternative solution, of course, is to recognize, and to persuade governments, that we are strictly on the linear portion of the graph of economic output versus scientific expenditure; that such expenditure is therefore an investment and not a cost; and thus that if countries such as the United Kingdom wish to enhance their economic performance, they should aim to increase the amount of investment in equipment, not decrease the number of scientists.

Douglas B. Kell

Edward Llwyd Building,
Institute of Biological Sciences,
University of Wales,
Aberystwyth SY23 3DA, UK
e-mail: dbk@aber.ac.uk

Sir — It is hard for a mere member of the research riff-raff to criticize someone as eminent as Sir Hermann Bondi, especially because, as a former chief scientific adviser

to the Ministry of Defence, he can be considered to be a field marshal of British science. It is important, however, that officers do not undermine the morale of the troops.

Some of us still have to earn our living and are still slogging away in the front line at the computer terminal or in the laboratory, and we scientific cannon fodder can really do without this kind of friendly fire.

Bondi is right to point out the inefficiency of the present research funding (or non-funding) system, but this is presumably something that evolved while he was a staff officer. Although I am sure that he is retired in name only, future cuts in research funding are unlikely to have any effect on his pension — something we other ranks will never get!

David McA. McKirdy

173A London Road,
Stoke-on-Trent,
Staffordshire ST4 5RW, UK
e-mail: gga33@keele.ac.uk

Credit where it's due

Sir — Friedrich Katscher is correct in stating that science has a long history of abuse of authorship for honour and glory¹.

Insulin, for instance, was discovered by Frederick Banting and Charles Best. Yet the Nobel prize for physiology or medicine was awarded in 1923 to Banting and John J. R. Macleod. Macleod was

merely the head of the physiology laboratory in Ontario where the research was done. Not only did he take no active part in the research, he was on vacation when the crucial experiments were done. Banting, in protest, shared his half of the prize money with Best; Macleod did the same with J. B. Collip, who had assisted with the purification and standardization of insulin².

It must be admitted, though, that there are gentlemen scientists, too. Klinefelter states that the syndrome named after him was the result of an unselfish action on the part of Dr Fuller Albright, who after they jointly described the syndrome, allowed Klinefelter to put his name first on the list of authors³.

I am collecting material on such eponymous misnomers in biomedical research, as well as examples of serendipitous discoveries in the field. I would welcome assistance from fellow scientists.

Sanjay A. Pai

Cellular Genetics—181,
University of Texas,
M D Anderson Cancer Center,
1515 Holcombe Blvd,
Houston, Texas 77030, USA
e-mail: sanjayapai@hotmail.com

1. Katscher, F. *Nature* 388, 320 (1997).

2. Odelberg, W., ed. *Nobel, the Man and his Prizes* (Nobel Foundation/Elsevier, New York, 1972).

3. Firkin, B. G. & Whitworth, J. A. *Dictionary of Medical Eponyms* (Parthenon Publishing Group, UK, 1987).

Climate science and insurance risk

Partnerships between insurance companies and climate scientists are a model for how academic science can provide value to business. These partnerships benefit both businesses and scientists, as well as the wider public.

**Anthony Michaels, David Malmquist,
Anthony Knap and Ann Close**

In August 1992, Hurricane Andrew ripped through southern Florida and Louisiana, causing an insured loss of \$16.5 billion and total damages of more than \$30 billion. This high level of loss occurred even though the strongest part of the storm missed the city of Miami. Payment of claims from Andrew made several insurance companies bankrupt and depleted much of the available worldwide insurance capital for natural catastrophes.

Andrew's aftermath led many insurers to reassess their catastrophe exposures and reinforced a growing awareness that traditional actuarial approaches to analysing catastrophe risks are inappropriate. This concern has been accentuated by a growing political debate in Europe (sparked in part by the environmental advocacy group Greenpeace) concerning poorly documented claims that global warming might increase the number of tropical cyclones. The losses from Andrew were followed by the creation of a new group of 'catastrophe' reinsurers in Bermuda. These companies specialize in covering the largest of aggregate-insured losses from individual catastrophes and the application of new technologies to assess the risk. By 1996, these new companies had more than \$5 billion in capital and accounted for a third of catastrophe reinsurance worldwide. As part of the search for new risk-analysis solutions to replace traditional actuarial approaches, a unique partnership between insurance companies and climatologists was formed in Bermuda in 1994. This is one model for how peer-reviewed scientific research can benefit businesses, and vice versa.

Applying science to risk transfer

Insurance represents a means to make financial risk more tractable by transferring it from individuals to large groups and institutions. Western industrialized countries use private insurance as the main way to transfer property risks that are too large to be handled by individuals to institutions large enough to cover the loss. Primary insurers write policies covering defined, unpredictable losses in exchange for a predictable premium. In this process, insurers build up large aggregate exposures to catastrophes which, by definition, affect large numbers of risks at once. Reinsurers, a subset of insurance companies, insure primary insurers against losses larger than the primary insurer wants to handle alone. They distribute this aggregate risk

over a larger capital base.

An understanding of hurricanes (and of climate in general) influences both traditional insurance practice and some of the newer methods of risk transfer that are under development. The premium that an insurer accepts must cover the expected loss costs, operating costs and the costs of capital or the anticipated profit (although some companies may lose money on their underwriting and make profits by successful investing). However, in any year, the insurer must be able to pay its claims, even if they exceed the amount collected in that year.

Insurers hold capital to guard against this uncertainty, and the amount of capital will be larger for riskier lines of business (like hurricane coverage). They must also pay for this capital or it will flow into another part of the economy that gives a greater risk-adjusted return on investment. Thus the riskier lines of business must pay larger returns. For example, a catastrophe reinsurer in Bermuda can yield as much as 20–30 per cent return on capital.

Better understanding of hurricane frequency on any timescale provides a more robust basis for assessing the likely losses associated with a catastrophe reinsurance contract. For traditional insurance practice, the timescale of contracts and business relationships ranges from one year to decades; these timescales of climate variability are most relevant to insurers in the traditional markets. When yearly or decadal losses differ greatly from the historical average, an understanding of climate may allow a more reasonable assessment of the risk and a rationale for a given policy's price or the allocation of capital.

Insurers who collect too little in premiums or who have too little capital stretched too far will fail when an increase in hurricane activity results in larger losses. This failure ultimately places the burden back on individuals or governments. A better understanding of climate processes will allow a more logical assessment and transfer of risk than is used at present. Insurers are regulated

...the largest tropical cyclones or earthquakes in a large city can generate insured losses of up to \$100 billion. This loss exceeds all the reinsurance capital...

and reinsurers share risks. These two aspects of the market require that new techniques for risk assessment, like the understanding of climate-induced changes in hurricane frequency, are publicly validated.

Insurance companies' interest in climate science is also beginning to focus on shorter-term forecasting skills as new market tools are developed. Insurers now realize that the reinsurance mechanism probably fails for the largest disasters because the pools of capital are too small. The capital in the property insurance industry is actually fairly modest: the total pool of catastrophe capital is about \$6–8 billion, whereas the capital pool for the general reinsurance industry is about \$30 billion and the capital that backs up the \$40,000 billion in property insured by primary companies is \$250 billion. Recent analyses suggest that the largest tropical cyclones or earthquakes in a large city can generate insured losses of up to \$100 billion. This loss exceeds all the reinsurance capital and would deplete much of the total pool of capital, bankrupting many companies. It is impossible to set aside more catastrophe capital without dramatically reducing the return on the capital or increasing the costs in a market that is already soft. One way to avoid this potential catastrophe is to transfer insured risk to another part of the global economy. The global capital markets are worth approximately \$20,000 billion, and the derivatives markets may exceed \$60,000 billion. Daily fluctuations in the total size of the capital markets exceed the largest possible insured loss. Thus, an event that depletes a third of the total insurance capital would be lost in the noise of the capital markets.

The challenge is to create tradeable financial products that allow people to trade on fluctuations in an index of insured risk. With a traditional actuarial approach, historical losses set the value of a risk, and risk values change very slowly. However, for climate-related risks, the likelihood of an event changes with time, and the many forecasting approaches allow for a diverse information stream about the future expected loss on timescales of hours to decades. Many new developments are required to create a capital market in catastrophe risk, some of which must come from peer-reviewed scientific research. Scientists need to develop new forecasting strategies on the entire spectrum of timescales and link these forecasts to the relevant hazard (hurricane landfall, for example). Insurers must believe that the strategies work (public validation via peer-reviewed publication) and understand them

AP

(education). However, traders in the financial market must also think that they know something that others do not or that they have an edge (perhaps by hiring climate scientists or their students to produce suitable applications of forecasts).

Mutual benefits

We all have much to gain from improved connections. Scientists need money for research; businesses have money and need research. As scientists understand more about the value of their work to a particular industry, they will find better ways to make it relevant. As businesses see a more realistic and wider view of the scientific world, they will provide opportunities and support for scientists, and perhaps even lobby governments. Further, in the case of insurance, there is an obvious benefit to society. The Western world uses insurance as its primary risk-transfer mechanism and hence needs insurers to succeed (despite our individual pain when insurance premiums fall due). If insurers fail, it is the home-owner and the taxpayer who lose. Thus, everyone benefits from research that puts the insurance industry on a more sensible risk-analysis footing via an improved ability to handle the extremes of climate.

The basic assumptions underpinning most of the insurance industry are violated by the laws of nature that apply to climate and tropical cyclones. Insurers traditionally assume that the average insured losses over a recent historical period accurately reflect future losses over some arbitrary future period (with some flexibility for trends). This assumption works well with automobile insurance and other widely distributed, independent risks, but is inappropriate for large catastrophic losses like hurricanes and earthquakes. Before 1986, no insured loss had exceeded \$1 billion. Hurricane Hugo in 1989 was five to six times higher than this limit. Hurricane Andrew and the Northridge earthquake caused losses an order of magnitude larger. Much of the increase in loss is due to increased development. Actuaries can modify their assumptions to recognize that the distribution of property changes over time (for example, that Florida's population has increased by 500% over the past 40 years). But this approach still assumes that the likelihood of storm landfalls in the next few years can be estimated from the average activity of the past 30–100 years.

By contrast, climate scientists understand that climate changes on all timescales and in all geographical regions; that the average activity in any arbitrary period of the past is not necessarily a good predictor of future activity; and that understanding the patterns in climate and the processes that control climate allows us to predict how future activity will differ from the average of the past. The types of patterns that climatologists study, for example the geographical linkages of the



Understanding risk — the devastation caused by Hurricane Andrew, which hit Florida City in August 1992, causing \$15 billion damage to the South Florida area alone.

coupled ocean–atmosphere system, are difficult for most insurers to understand and hence to incorporate into business decisions, because they do not have the necessary science background and much of the information is fairly difficult for them to synthesize from the scientific literature.

The first challenge of creating a link between information and understanding is to identify the role of peer-reviewed science in providing value to a specific business community. Businesses are relatively good at finding proprietary information that helps them to beat their competitors. But proprietary information is the antithesis of the academic approach, as it robs the wider scientific community of expertise, understanding and data. In their normal search for proprietary advantage, most companies do not see any value in the shared body of understanding that is the heart of academic pursuit. Yet sharing risks and the public regulation of insurers provides the background for identifying value in academic approaches. Defining the 'value' of a piece of public science can sometimes be as simple as asking a company exactly what it needs a competitor to know so that the competitor's ignorance does not force everyone to make a stupid decision.

The insurance industry is a logical initial target for developing academic–business interactions that define the value of certain kinds of public information. First, the industry is tightly bound by many traditional actuarial methods of risk analysis. Thus, introduction of a new analytical approach

requires a public validation (peer-review) before it is acceptable. Second, primary insurers in the United States and elsewhere are tightly regulated by state governments, so regulators (and the electorate) need to be convinced of the reason for changes in company pricing and other behaviour. Third, even in the less-regulated reinsurance industry, individual companies often share risks with others. If these other companies do not understand a novel method of risk analysis, they may not be willing to share the risk. Hence, certain kinds of public scientific information have a real 'value'. Fourth, the introduction of capital market strategies requires a broad acceptance of new forecasting strategies to allow both the buyer and seller to feel that they are on even ground.

Climate events relevant to insurers

What then are the scientific questions of interest to insurers? Through the Bermuda-based 'Risk Prediction Initiative', a dozen global insurance companies and an international group of scientists have been discussing this intersection between catastrophe exposures and the study of tropical cyclones and climate. Insurance companies are interested in the following parameters.

First, they care about events that cause insured losses. They care about hurricanes and typhoons that make landfall in developed, insured countries, but can ignore hurricanes that never cause damage on land. They worry less about hurricanes that make landfall in countries with low levels of insurance, even though these countries may still



have significant property damage and loss of life. They care little about droughts unless they lead to large fires. They care about floods in Europe, but not in the United States (flood insurance in the United States is covered by the federal government).

Second, the timescale of insurance contracts, insurer–client relationships and the timescale within which a company can take acceptable actions to modify its risk position all determine the climate timescales that will be relevant. Reinsurers tend to write contracts for events within a single year, and develop continuous relationships with client insurers over many years. Both insurers and reinsurers have the ability to transfer some risk elsewhere through reinsurance and retrocession contracts. Thus, research on the tracks of individual storms may be valuable to the evacuation of coastal populations, but is less relevant to the contract timescale of insurers. However, as insurers develop the capital market products discussed earlier, they will become more interested in the shorter timescales characteristic of the commodities markets. Climate variability on the multi-annual timescales currently relevant to insurers is dominated by the El Niño Southern Oscillation, a fluctuation in the distribution of heat in the equatorial Pacific.

Variability on timescales of five to 20 years is also important. The decadal climate cycles now being recognized in the Pacific, Atlantic and Southern oceans seem to provide a rationale for understanding some of the longer-term fluctuations in the frequency of tropical cyclone landfall. The timescale of global warming is too long for most business decisions from a risk-analysis standpoint, but may still influence companies'

decision-making processes from the political perspective. However, the highly politicized 'greenhouse' debate is sprinkled with claims that global warming will increase the frequency of intense tropical cyclones. This contention is not supported by widespread consensus among climate specialists; indeed many studies indicate that the opposite may be true. Yet an understanding of the scientific basis of the claims is crucial.

Third, insurers and reinsurers care about the accurate definition of catastrophe risks, the distribution of risk in time and space, and how these distributions change with time. Reinsurers are in the business of taking risks, particularly risks priced to generate a profit. Increases in the frequency of hurricane landfall may logically lead to higher prices or restructured contracts, and decreases in the frequency may lead to lower prices of reinsurance. Prices also respond to events. A catastrophe removes capital from the industry and drives prices up, whereas prices sharply decline during periods of fewer losses. This cyclic pattern occurs in the absence of information about future risk, often leading to prices that differ significantly from the expected losses. Such instability is not in the interests of the insured, who are relying on the good judgement of the insurance industry to protect them from these events. More accurate risk assessment should lead to stability and more comprehensible pricing.

Finally, insurance companies need information in a form that allows them to make a decision. This is an area that we think needs much more development. Over the past decade, risk-analysis software packages have come on the market that forecast future losses and give the probability of a loss exceeding a given size. These packages use a layered analysis that includes a probabilistic model of the hazards, the likely damage from each hazard, and the financial implications of each event. Current generations of these models use some published scientific research, and each has been exposed to limited peer-review during its development. However, all these models assume that the historical storm data sets represent the future frequencies of storms, and they are all fundamentally proprietary for many important features (particularly in how they transfer uncertainty through the models). These models represent a crucial change in how insurers look at risk from episodic catastrophes, and they represent a structure for transferring new climate research into decisions by underwriters.

Where next?

Determining the value of science in a business context requires that scientists appreciate the role of science in a business decision-making process and become familiar with the forces outside their expertise that drive business decisions. This broader understanding of business occurs regularly in sci-

entific fields that produce discrete products, such as inventions or drugs that have defined market mechanisms (for example, patents) for specifying ownership. The transfer of knowledge to guide decisions is more difficult. The decision process is at the heart of the proprietary elements of a company, and knowledge is more difficult to patent or copyright than a discrete product. Nevertheless, careful integration of science into decision tools is crucial if businesses are to accept the value of relevant scientific research.

One model for how academic science can provide value to the business community is the Risk Prediction Initiative mentioned above, formed specifically to support and explain research on the frequency and severity of tropical cyclone landfalls. The initiative runs workshops for insurance companies and scientists to integrate science into the decision processes in insurance. Research projects are first scientifically peer-reviewed and then evaluated by a group of insurance companies to determine their relevance. Publication is a condition of continued funding. Many different insurance companies participate in this programme: primary insurers, reinsurers and catastrophe reinsurers, brokers, modellers, and technical service providers (see acknowledgements). The initiative is just one example of how to create links between science and business that will ultimately benefit everyone.

As this programme has grown, it has provided insights into other areas where the links between science and the decision process could provide value. In many debates about pollution, global warming, product liability and public health, advocacy groups and businesses use science to support predetermined positions. Better links between science and the business-decision process would encourage alternative solutions that would be better for business and society. This is precisely the role that science should play. If scientists are active in the creation of decision tools, it increases the value of the science without compromising the scientists' perception of their own objectivity. If scientists do not themselves work to make science relevant wherever possible, no-one else will. □

*Anthony F. Michaels and Ann Close are at the USC Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, California 90089-0371, USA.
e-mail: tony@usc.edu*

David Malmquist and Anthony Knap are at the Bermuda Biological Station for Research, Ferry Reach, Bermuda GE 01.

Acknowledgements. The Risk Prediction Initiative (RPI) was founded by scientists at the Bermuda Biological Station for Research and has received significant support from that institution. RPI is currently supported by Mid-Ocean Re, CAT, Renaissance Re, Terra Nova Insurance, Center Re-Zurich, ACE, X.L. Insurance Co., General Re, Employers Re, American International Group, Tillinghast Towers-Perrin and in the past by Chubb group and AON Corp. We thank all participants and sponsors of RPI. Our Web pages are at <http://wrigley.usc.edu/> and <http://www.bbsr.edu/>

Efficient fish not faint-hearted

Elizabeth Brainerd

The blood circulation systems of certain air-breathing fishes and most reptiles appear to be primitive and inefficient. But there may well be good reason that they are designed the way they are, and the principles concerned could inform clinicians who treat heart disease.

A primary function of blood circulation is to provide oxygenated blood to the tissues of the body. Why, then, in some air-breathing fishes, does oxygenated blood flow from the lung to the heart and then to the gills, rather than flowing directly to oxygenate the body? And why do the hearts of most reptiles retain an only partially divided ventricle, given that mixing venous and arterial blood reduces oxygen delivery to the body? Writing in the August issue of *Paleobiology*, Colleen Farmer¹ proposes an intriguing function for these puzzling circulatory patterns — that they provide oxygenated blood to the heart muscle itself.

Gar and bowfin, two North American freshwater fishes, possess both gills and a lung, and are therefore able to extract oxygen from both water and air. In these fishes, oxygenated blood drains from the lung directly back to the heart, where it mixes with venous blood and is then pumped through the gills before flowing out to the body (Fig. 1a). Comparative physiologists have traditionally viewed such systems as primitive and inefficient because oxygenated blood from the lung is diluted with deoxygenated blood before being pumped to the body. Furthermore, some fishes that have this system could actually lose oxygen through the gills if the surrounding water is very deoxygenated. Farmer proposes, however, that this circulatory design serves the function of providing highly oxygenated blood to the heart muscle (myocardium) itself, which may help to explain the persistence of an apparently inefficient circulatory design through evolutionary time.

In most fishes, amphibians and reptiles, the ventricle of the heart lacks a dedicated coronary circulation. Much of the heart muscle is arranged in a loose, spongy matrix that receives oxygen directly from the blood as it is pumped through the heart chambers. The lung of an air-breathing fish provides a source of highly oxygenated blood that flows directly back to the heart (Fig. 1a). Oxygen supplied by the lung may be particularly important for the heart when an air-breathing fish exercises, during which the swimming muscles increase their oxygen extraction from the arterial blood. Thus the venous blood returning to the heart contains even less oxygen than it does when the fish is at

rest, precisely at a time when the heart is working harder and needs more oxygen.

In birds and mammals, the systemic (body) circulation is completely divided from the pulmonary circulation (Fig. 1c). This division has traditionally been viewed as the most efficient circulatory design because venous and arterial blood do not mix, and so only maximally oxygenated blood is sent to the body. In turtles and squamate reptiles (lizards and snakes), however, the ventricle is only partially divided. Blood-flow patterns in the heart generally keep arterial and venous blood separate, but some mixing can occur (Fig. 1b). As with the air-breathing fishes, this incomplete division of the heart was historically seen as a primitive, intermediate system on its way to becoming the more efficient mammalian design.

More recently, physiologists have recognized that the partially divided circulation of turtles and squamates may be an adaptation for the intermittent breathing pattern of such animals²⁻⁴. These studies emphasize the

shunting of blood from the right to the left ventricle, which diverts blood away from the lungs when they are not in use. But a perplexing left-to-right shunt of blood has also been measured during exercise in reptiles, in which part of the oxygenated blood from the lungs is sent from the left to the right ventricle and thence back to the lungs⁵.

Farmer proposes that the left-to-right shunt provides oxygenated blood to the spongy myocardium of the right ventricle. Without this intracardiac shunt, the right side of the heart would receive only deoxygenated, venous blood, and could suffer from lack of oxygen during activity. In birds and mammals, this problem has been circumvented by the evolution of coronary circulation that feeds oxygenated blood from the aorta back to the left and right sides of the heart. Coronary circulation has also allowed the evolution of compact myocardium, further reducing the uptake of oxygen directly from the blood flowing through the lumen of the heart.

Medical researchers seeking treatments for blocked coronary arteries are now becoming interested in the spongy myocardium of fishes and reptiles⁶. Surgeons have begun using lasers to punch holes in the left ventricle to permit some oxygenation directly from the luminal blood, a technique known as transmural revascularization^{6,7}. Farmer's ideas about the oxygenation of fish hearts could inspire other treatments. Perhaps it would be feasible to implant a device, analogous to a fish lung, that would inject a small amount of a highly oxygenated solution directly into the coronary arteries, either as a temporary remedy during an acute heart

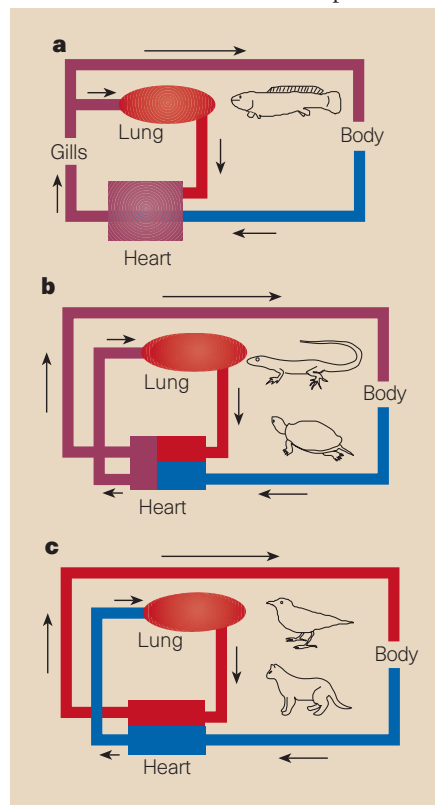


Figure 1 Simplified circulatory-system diagrams of selected vertebrates. a, The circulatory system of certain air-breathing fishes is relatively inefficient at delivering oxygen to the body because blood flows from the lung back to the heart where it mixes with venous blood before being pumped to the body. b, Turtles and squamates (lizards and snakes) are able to maintain functional separation of arterial and venous blood, but they often shunt blood across a connection between the ventricles, thus mixing the flows. c, Birds and mammals have coronary arteries that feed oxygenated blood to both the left and right ventricles, and so have been able to evolve completely divided hearts. The myocardial oxygenation theory, proposed by Farmer¹ and discussed here, proposes a function for the apparently inefficient circulatory designs depicted in a and b. The hearts of most fishes, amphibians and reptiles lack coronary circulation, and instead obtain oxygen directly from the blood flowing through the chambers. Farmer suggests that oxygenated blood flowing from the lung to the heart of air-breathing fishes provides oxygen directly to the heart muscle, and that shunting from the left to the right side of the heart in reptiles provides oxygen to the right ventricle. Fully oxygenated blood is red; deoxygenated blood, blue; partially oxygenated blood, purple.

attack, or possibly as a longer-term alternative to bypassing or clearing partially blocked coronary arteries.

Farmer formed her ideas about the oxygenation of spongy myocardium through the study of circulatory anatomy and physiology in a variety of non-mammalian vertebrates. Her theories will be testable through physiological studies of circulation, blood-gas concentrations and cardiac function during exercise in fishes, amphibians and reptiles. She also proposes another hypothesis, one that will be much tougher to test, in which lungs evolved in the common ancestor of fishes and tetrapods primarily to provide oxygen to the heart. According to the current view, lungs evolved mainly to increase the oxygen available to the whole body, and thus permitted the early bony fishes to invade poorly oxygenated aquatic environments, such as swamps^{8,9}. Other views of lung evolution have also emphasized the importance of lungs in regulating the buoyancy of aquatic animals with dense, dermal armour^{9,10}.

Farmer marshals evidence for the plausibility of her evolutionary theories from palaeontology, ecology and physiology. It is difficult, however, to argue that a feature of an organism evolved for only one of several functions that the feature may serve today, without some evidence from an evolution-

ary analysis that one function evolved before another. It may not even be very useful to attempt to determine whether systemic oxygenation, myocardial oxygenation or buoyancy control was the most important selective factor in the evolution of lungs, because we can be certain that their evolution had inescapable consequences for all three.

Nonetheless, Farmer has done a great service in bringing this third important function of lungs to our attention, and in providing an elegant explanation for the evolutionary persistence of some seemingly inefficient circulatory designs. □

Elizabeth Brainerd is in the Biology Department, and the Organismic and Evolutionary Biology Program, University of Massachusetts, Amherst, Massachusetts 01003, USA.

e-mail: brainerd@bio.umass.edu

1. Farmer, C. *Paleobiology* 23, 358–372 (1997).
2. White, F. N. in *Biology of the Reptilia* (eds Dawson, W. R. & Gans, C.) 275–334 (Academic, New York, 1976).
3. Johansen, K. & Burggren, W. W. in *Hearts and Heart-Like Organs* (ed. Bourne, G.) 61–117 (Academic, New York, 1980).
4. Hicks, J. W. & Wang, T. J. *Exp. Zool.* 275, 204–216 (1996).
5. Wang, T., Krosniunas, E. H. & Hicks, J. W. *Am. Zool.* 37, 12–22 (1997).
6. Kohmoto, T. et al. *Circulation* 95, 1585–1591 (1997).
7. Horvath, K. A., Manning, F., Cummings, N., Shernan, S. K. & Cohn, L. H. J. *Thorac. Cardiovasc. Surg.* 111, 1047–1053 (1996).
8. Gans, C. *Forma et Functio* 3, 61–104 (1970).
9. Liem, K. F. *Am. Zool.* 28, 739–759 (1988).
10. Graham, J. B. *Air-Breathing Fishes: Evolution, Diversity, and Adaptation* (Academic, New York, 1997).

Particle physics

Exotic meson plays hard-to-get

Frank Close

According to the popular media of the United States and now Europe, physicists have found the 'glue that holds life, the Universe and everything' together. The Web page of the E852 Collaboration at the Brookhaven National Laboratory in New York proclaims "we are in the news". In more measured language, in a paper published in *Physical Review Letters* this month¹, they have "evidence for exotic meson production"; and at the Hadron97 conference, held at Brookhaven on 26–31 August, these claims were critically evaluated. Something definitely seems to be going on, but precisely what is it?

It is well established that the seeds of protons, neutrons and all strongly interacting particles ('hadrons') are quarks. According to the theory of quantum chromodynamics, which is part of the Standard Model of particle physics, quarks are gripped together by the exchange of gluons. Although this is established at high energies, the dynamics of gluons at low energies, where the forces become strong and the proton and other hadrons are formed, are obscure.

Hundreds of short-lived hadrons have been found during the past 50 years. They are 'resonances', where the quarks are excited into states of higher energy; but in none of

these has the glue played other than a passive role. Yet according to quantum mechanics, the gluon degrees of freedom also should be excitable in the presence of quarks, forming 'hybrid' states in which both quarks and gluons are active components. The energy of the excitation adds mass to the system, and also allows characteristic correlations between various properties (such as spin and the behaviour of the wavefunction under certain transformations) that would normally be impossible.

In the rather primitive theoretical models of the early 1980s, we noticed that the lightest of these 'exotic' states could have a mass of around 1.5 giga-electron volts (GeV) — some 50 per cent more than a proton^{2–4}. In the intervening 15 years theory has matured, and the mass of the lightest exotic is expected to be around 1.8 GeV, with a lifetime of 10^{-23} s and most probably decaying into a pion accompanied by more massive particles such as the η or f_1 (refs 5–8). And there have been a few hints of experimental detections.

In 1988 the GAMS Collaboration⁹ reported a particle at 1.4 GeV, but this was quickly criticized and has been dismissed¹⁰. Then in 1993 two claims emerged, one from Japan's KEK laboratory¹¹ and the other from

the VES collaboration at Serpukhov¹². There were some differences of detail between this pair, and together with the emerging doubts about the earlier GAMS claim, there was a 'once bitten twice shy' reaction in the community. But the new publication from Brookhaven, and preliminary news from the Crystal Barrel collaboration at CERN, has led to a real resurgence of interest.

The VES experiment¹² used beams of pions with 37 GeV c^{-1} incident momentum to produce their exotic object, which was revealed through its decay into a π -meson and an η -meson. The new Brookhaven experiment¹ uses a less energetic incident pion beam, but the results are very similar: they see a π and an η , and the inferred mass and lifetime of the implied exotic particle are not inconsistent with the VES observations. So reports of a 'discovery' may be less than fair to the earlier Russian experiment; ironically so, as some members of the original VES team are collaborators in the BNL experiment.

The experiment does not see a clear resonance, but rather a spatial asymmetry in the distribution of the η and π together with a distinctive dependence on energy of a certain quantum-mechanical phase. These phenomena are interpreted as evidence for a quantum-mechanical interference between a well-known resonance (the 'a2') and the conjectured exotic particle. The effect is subtle, and there was some concern over whether a small error in quantifying the intensity of the a2 signal could have misled the researchers. However, confidence in the result was increased by a report at Hadron97 that the Crystal Barrel collaboration at CERN sees similar phenomena when they study the π and η produced in annihilations between antiprotons and matter (K. Peters, Univ. Bochum; E. Klempt, Univ. Bonn). Reassuringly, the production mechanism, research team and systematic uncertainties all differ from those in the BNL and VES experiments.

Were this the whole story, it would be natural to conclude that the long-sought evidence of an 'exotic' had been found. But the fuller panoply of evidence presented at Hadron97 has created an enigma.

In a related experiment at Brookhaven, attempts have been made to find the same exotic particle decaying into something other than a π and an η . A signal is seen, but now at around 1.6 rather than 1.4 GeV. In investigations looking at yet other predicted decay paths, an experiment at Brookhaven and an independent one in Russia see a signal at around 1.8 GeV.

Have we searched for over ten years suddenly to see three exotics come along at once, like the proverbial London buses? This seems unlikely. There may be a single, genuine, exotic resonance around 1.6 to 1.8 GeV, and then established quantum effects

associated with the presence of several different possible decay channels could cause activity at different energies in different specific channels.

Theorists will now examine whether this interpretation fits the phenomena. There are many more data from Brookhaven to be analysed, and together with the continuing analysis of CERN and VES data, these should begin to answer the conundrum. □

Frank Close is in the Theory Division,
Rutherford-Appleton Laboratory, Chilton,
Didcot OX11 0QX, UK.

1. Thompson, D. R. *et al.* *Phys. Rev. Lett.* **79**, 1630–1633 (1997).
2. Barnes, T. & Close, F. E. *Phys. Lett. B* **116**, 365–368 (1982).
3. Barnes, T., Close, F. E. & de Viron, F. *Nucl. Phys. B* **224**, 241–264 (1983).
4. Chanowitz, M. & Sharpe, S. *Nucl. Phys. B* **222**, 211–244 (1983).
5. Lacock, P., Michael, C., Biyle, P. & Rowland, P. *Phys. Lett. B* **401**, 308–312 (1997).
6. MILC Collaboration hep-lat/9707008 at xxx.lanl.gov
7. Isgur, N., Kokoski, R. & Paton, J. *Phys. Rev. Lett.* **54**, 907–911 (1985).
8. Close, F. E. & Page, P. R. *Nucl. Phys. B* **443**, 233–254 (1995).
9. GAMS Collaboration *Phys. Lett. B* **205**, 397–402 (1988).
10. Prokoshkin, Y. D. & Sadoviski, S. A. *Phys. Atoms Nuclei* **58**, 606 (1995).
11. Aoyagi, H. *et al.* *Phys. Lett. B* **314**, 246–252 (1993).
12. Beladidze, G. M. *et al.* *Phys. Lett. B* **313**, 276–283 (1993).

Chromatin structure

The nucleosome core all wrapped up

Daniela Rhodes

For many years, biologists have been trying to understand how DNA is packaged into chromosomes. A breakthrough came in the 1970s when it was established that all eukaryotic chromosomes consist of a regularly repeating protein–DNA complex called the nucleosome¹. Each nucleosome consists of a protein octamer, made up of two copies each of histones H2A, H2B, H3 and H4, which, together with the fifth histone, H1, organizes about 200 base pairs (bp) of DNA. Further organization involves the assembly of nucleosomes into higher-order chromatin structures.

On page 251 of this issue, Luger *et al.*² report the much awaited high-resolution structure of the nucleosome core particle. This landmark structure confirms many of the expected features of the nucleosome core. But it also reveals some surprises. Importantly, it provides a structural basis for interpreting the many molecular mechanisms that have evolved to regulate access to the DNA in chromatin.

The low-resolution (7-Å) structure of the nucleosome core particle was determined in 1984, and it revealed that the histone octamer forms a helical ramp around which is wrapped 1.7 turns of a left-handed DNA superhelix³. Now, the higher-resolution structure (2.8 Å) clearly shows how the nucleosome core is assembled. Each histone consists of a structured, three-helix domain called the histone fold, as well as two unstructured tails. The histone-fold domains pack through a handshake-like arrangement to form the heterodimers H2A–H2B and H3–H4. These oligomerize further to form the histone octamer⁴. Interestingly, the histone-fold motif has turned up in a number of transcriptional coactivators and repressors, providing an evolutionary link between chromatin and the transcriptional machinery⁵.

In the structure solved by Luger *et al.*, we can see — at near atomic detail — how DNA is bound and organized by the histone core

(Fig. 1). The authors used DNA of defined sequence (146 bp), 121 bp of which is organized by the histone-fold domains. Where the DNA enters and leaves the nucleosome, it is bound by extensions of the H3 histone fold. Each of the heterodimers binds about 30 bp of DNA. The DNA is contacted at 10-bp intervals as the minor groove faces the protein. Many contacts to the phosphate backbone of the DNA are made from the main-chain atoms of the protein. This type of contact is needed for the DNA to be held rigidly but, at the same time, for the interactions to be nonspecific so that the many different DNA sequences in a genome can be accommodated. In addition, arginine side chains penetrate all 14 of the minor grooves that face the protein core.

The path of the DNA superhelix is not uniform — it is distorted through bends at several positions. Each H3–H4 and H2A–H2B pair generates these bends, but the two types of dimers do it differently. This distortion matches that seen in the 7-Å

nucleosome core structure containing mixed-sequence DNA³.

The average number of base pairs per helical turn of DNA — the helical periodicity of DNA — has a value of 10.6 bp in solution, but Luger *et al.* find that, as expected⁶, the helical periodicity around the nucleosome core is 10.2 bp. Together with the distortion in the superhelix, this provides a structural explanation for why nucleosomes take up preferential positions with respect to DNA sequence^{7,8}. A value of 10.2 bp allows the minor (and major) grooves from neighbouring turns of the DNA superhelix to line up, forming channels through which the histone tails can pass. This also leaves the major grooves accessible, enabling them to participate in cellular processes, acting, for example, as DNA integration-hotspots⁹.

A remarkable feature of the new nucleosome core structure² is the ordered nature of the histone amino-terminal tails, and the involvement of these tails in nucleosome–nucleosome interactions (Fig. 2, overleaf). Although they are very long, the tails show no secondary structure and they are visible for about one-third of their length. The amino-terminal tails of both H3 and H2B pass through minor-groove channels, so there is a protruding tail every 20 bp. Both of H3's tails extend considerably beyond their exit points at the nucleosomal DNA surface, and they are fixed in this orientation by nucleosome–nucleosome contacts. One of the protruding H4 tails makes many contacts with the face of an H2A–H2B dimer from a neighbouring nucleosome core. Because crystals were obtained under near-physiological conditions, what is seen in the structure probably also applies in solution.

During transcription, the chromatin must unfold to allow the transcriptional machinery access to the DNA. Increased acetylation of histone tails is associated with

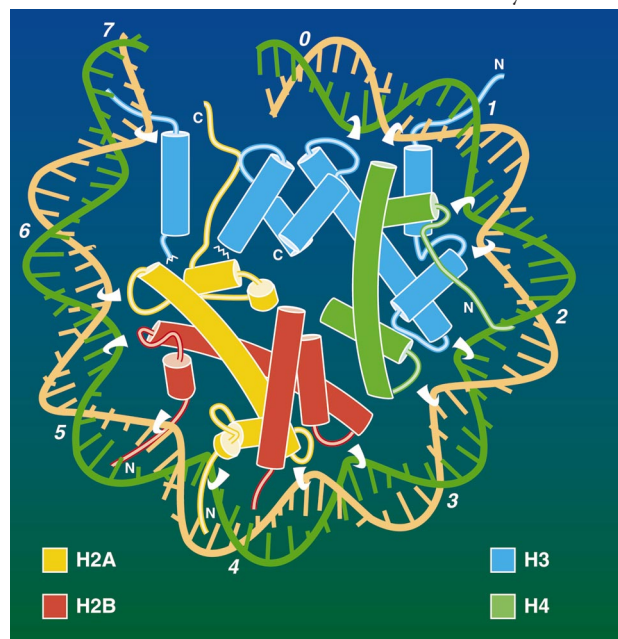


Figure 1 Luger *et al.*² have solved the structure of the nucleosome core to a resolution of 2.8 Å; half of the structure is shown. One turn (73 bp) of the DNA superhelix is visible, and the view is down the superhelical axis. The different histones are indicated by different colours, and helical regions in the proteins are represented by cylinders. The centre of the nucleosome (the dyad position) is indicated by 0 above the DNA superhelix, and it terminates at site 7. Points of protein–DNA contacts are indicated by white hooks.



Figure 2 The structured tails of histones. This view of the structure of the complete histone octamer in the absence of DNA was created by removing the DNA from the nucleosome core image (Fig. 1 on pages 252–253). The view is down the superhelix axis, and the different histones are colour-coded as in Fig. 1. (Courtesy of Tim Richmond.)

transcriptional activity, and underacetylation has been linked with repression¹⁰. The discovery that both acetyltransferases and deacetylases form complexes with certain transcription factors, provides a mechanism by which the appropriate region of chromatin can be targeted for 'remodelling'¹¹. The prevailing view in the literature is that the reduction in net charge offered by acetylation releases the histone tails from the surface of the nucleosome, leading to a conformational change in the nucleosome itself¹². But the structure of the nucleosome core particle — in which the amino-terminal tails make nucleosome–nucleosome interactions — puts this interpretation into question. Instead, it suggests that acetylation promotes the disruption of the higher-order structure of chromatin, consistent with experimental data¹³. Furthermore, because histone tails contain many targets for acetylation, the disruption can be controlled by the acetylation of certain specific residues but not others¹⁰.

The structure reported by Luger and colleagues also gives an insight into transcriptional repression of telomeres. The region of the H4 amino-terminal tail that is required for repression is the binding site for the silencing factor SIR3 (ref. 14). In the structure, this region interacts with the H2A–H2B dimer in an adjacent nucleosome. This suggests that binding of SIR3 would require the disruption of nucleosome–nucleosome interactions.

Why did it take so long to solve the high-resolution crystal structure of the nucleosome core, given that it was first crystallized in 1977¹⁵, and that the 7-Å structure was solved seven years later³? To obtain crystals that diffracted to high resolution, Luger *et al.* reconstituted the nucleosome core using a suitable defined-sequence DNA, and assem-

bling a histone octamer from recombinant proteins. This is a real *tour de force* of biochemistry. Furthermore, the crystals diffracted weakly, so data collection had to wait for developments in crystallographic techniques, such as intense X-ray synchrotron sources and cooling of crystals to the temperature of liquid nitrogen, to reduce radiation damage.

Now, the high-resolution nucleosome core structure provides a much-needed basis for the interpretation of genetic and biochemical data, and it will undoubtedly help us in designing better experiments to study chromatin 'remodelling'. The stage is set for the next challenge — understanding how a string of nucleosomes folds into a higher-order structure of chromatin, and what the role of histone H1 might be. □

Daniela Rhodes is in the MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

e-mail: rhodes@mrc-lmb.cam.ac.uk

- Kornberg, R. D. *Annu. Rev. Biochem.* **46**, 931–954 (1977).
- Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. *Nature* **389**, 251–260 (1997).
- Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. *Nature* **311**, 532–537 (1984).
- Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E. & Moudrianakis, E. N. *Proc. Natl Acad. Sci. USA* **88**, 10148–10152 (1991).
- Burley, S. K., Xie, X., Clark, K. L. & Shu, F. *Curr. Opin. Struct. Biol.* **7**, 94–102 (1997).
- Prunell, A. *et al. Science* **204**, 855–858 (1979).
- Travers, A. A. *Trends Biochem. Sci.* **12**, 108–112 (1987).
- Rhodes, D. & Klug, A. *Nature* **286**, 573–578 (1980).
- Pryciak, P. M. & Varmus, H. E. *Cell* **69**, 769–780 (1992).
- Turner, B. M. *Cell* **75**, 5–8 (1993).
- Tsukiyama, T. & Wu, C. *Curr. Opin. Genet. Dev.* **7**, 182–191 (1997).
- Wolffe, A. P. *Nature* **387**, 16–17 (1997).
- von Holde, K. & Zlatanova, J. *Proc. Natl Acad. Sci. USA* **93**, 10548–10555 (1996).
- Hecht, A., Laroche, T., Strahl-Bolsinger, S., Gasser, S. M. & Grunstein, M. *Cell* **80**, 583–592 (1995).
- Finch, J. T. *et al. Nature* **269**, 29–36 (1977).



100 YEARS AGO

The energetic Secretary of the Society for the Protection of Birds has issued another letter of protest against the wanton destruction of birds to supply ladies with wings and feathers and stuffed skins for their bonnets. The Society is unceasing in its efforts to open the eyes of the gentler sex to the cruelty often practised in procuring plumes, and to the gradual extermination of many beautiful and beneficial birds. We regret to think, however, that such trifling matters do not disturb the minds of the majority of women when they choose their millinery. Present effect is to them the sole criterion of the value of a bonnet, and how the effect is produced they complacently leave others to inquire.

From *Nature* 16 September 1897.

50 YEARS AGO

'Gas-Turbine Propulsion in a Naval Vessel' – Messrs. Metropolitan-Vickers Electrical Co., Ltd., Trafford Park, Manchester, have installed gas-turbine propulsion equipment in the experimental naval craft, *M.G.B. 2009*, the trials of which will take place in the near future. The Company claims that this is the first naval vessel ever to be propelled by a gas-turbine. The characteristics of the simple-cycle gas-turbine include low overall specific weight and size with rapid starting, and these qualities make it very suitable for light vessels where high speeds may be required for limited periods and at short notice. In *M.G.B. 2009*, normal cruising and astern power is provided by two 1,250 B.H.P. 2,400 r.p.m. Packard internal combustion reciprocating engines, each engine driving its own propeller through a reduction gear. Maximum ahead power is obtained by bringing into operation a completely independent Metropolitan-Vickers gas-turbine of 2,500 B.H.P., which drives a third propeller through speed-reduction gearing to supplement the power of the reciprocating engines.

From *Nature* 20 September 1947.

Many more abstracts like these can be found in *A Bedside Nature: Genius and Eccentricity in Science, 1869–1953*, a 266-page book edited by Walter Gratzer. Contact David Plant. e-mail: subscriptions@nature.com

Origins of life

A left-handed Solar System?

Christopher F. Chyba

How and why did life on Earth come to use almost exclusively laevorotatory, or left-handed amino acids (L-enantiomers), rather than their mirror-image dextrorotatory, or right-handed forms (D-enantiomers)? The paper by Engel and Macko on page 265 of this issue¹ re-examines the Murchison meteorite, and reinforces the surprising result that some process favouring L-enantiomers seems to have operated before the origin of life on Earth, and probably before the formation of the Solar System. If this is correct, our Solar System may have formed with a built-in bias for left-handed amino acids.

An excess of L-enantiomers is an intriguing result, as it suggests that the handedness, or chirality, of terrestrial biology might have an extraterrestrial cause. But an L-excess indigenous to the meteorite is difficult to prove, as it may instead be due to contamination by the terrestrial biosphere of an originally racemic mixture. (In a racemic mixture, L- and D-enantiomers are present in equal abundance, as in typical laboratory chemical syntheses of amino acids.)

When amino acids were first discovered in the Murchison meteorite², apparent excesses of L-enantiomers ranging from 0 to 20% were found. However, the conservative interpretation of these results was that contamination was to blame, especially as non-protein amino acids were reported to show no enantiomeric excess³. ('Protein' amino acids are of the type used by terrestrial life in proteins, and are therefore the most likely to suffer from biological contamination.) A later study⁴, claiming indigenous enantiomeric excesses in Murchison of up to 70%, was criticized on similar grounds⁵.

This controversy led Pillinger⁶ to emphasize the need for a criterion independent of enantiomeric excess to judge the authenticity of meteoritic amino acids. He suggested determining the isotopic composition of individual enantiomers, because terrestrial stable isotope ratios differ from those in meteorites.

Engel and Macko¹ appear to have achieved this. These authors had previously⁷ used the carbon isotope ratio $^{13}\text{C}/^{12}\text{C}$ of alanine in Murchison to argue for an excess of L-alanine over D-alanine, but reservations about these measurements persisted⁸. In their present letter¹, they instead examine the $^{15}\text{N}/^{14}\text{N}$ ratio of individual Murchison amino-acid enantiomers. They find L-enantiomer excesses in the amino acids alanine and glutamic acid — both protein amino acids — of over 30% and 50%, respectively, with $^{15}\text{N}/^{14}\text{N}$ ratios that are too high to include much terrestrial material. The contradiction between this and an earlier report² that alanine and glutamic acid have L-enantiomer excesses of about 0 and 10%, respectively, suggests that the Murchison meteorite is very heterogeneous.

Another way to establish an extraterrestrial origin is to use amino acids that are extremely rare in the terrestrial biosphere. This was used by Zhao and Bada⁹ to argue that two amino acids found above and below the 65-million-year-old Cretaceous/Tertiary

boundary layer in Stevns Klint, Denmark, were not terrestrial contaminants. The same criterion was recently applied by Cronin and Pizzarello¹⁰ to the Murchison meteorite. These authors discovered enantiomeric excesses of up to 9% in apparently non-biological amino acids. So it appears that two very different approaches in two laboratories confirm enantiomeric excesses in the Murchison meteorite.

The same isotope ratios that distinguish Murchison amino acids from their terrestrial counterparts imply that these molecules or their chemical precursors originated in interstellar clouds^{1,8}. Why should chemistry in such a cloud have a bias towards one-handedness? One possibility¹⁰ is that an enantiomeric preference may have been imposed on the cloud out of which our Solar System formed by circularly polarized light, perhaps synchrotron radiation from a neutron star¹¹. If a substantial fraction of the organic inventory of early Earth was then derived from comets and asteroids¹², the synchrotron-radiation hypothesis would connect terrestrial biochemistry with the extreme physics of collapsed stars.

If the excess was indeed established by some process specific to the molecular cloud out of which our Solar System formed, similar enantiomeric excesses should be found in other organic-rich meteorites and in comets. Examination of other carbonaceous chondrites could therefore test this hypothesis. Unless there is a more widespread process operating, other solar systems could have been formed with a preference for D-amino acids, or with no preference at all — with

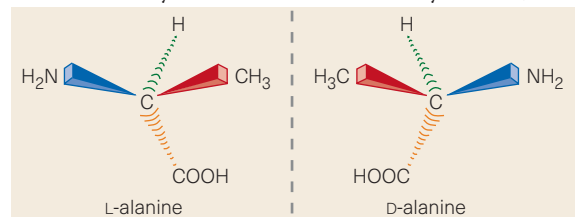
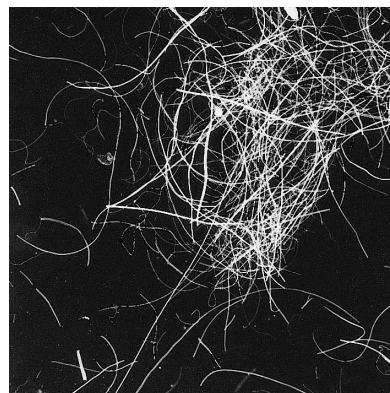
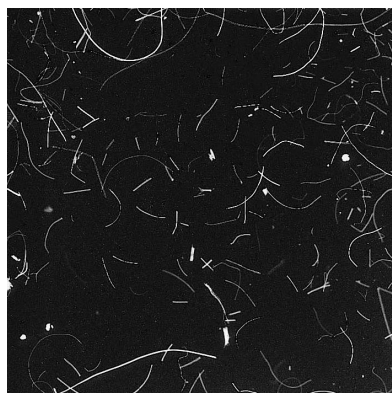


Figure 1 Left- and right-handed versions of the amino acid alanine. A meteoritic excess of L-alanine has now been shown not to be from terrestrial contamination, so life's chemical left-handedness may be extraterrestrial.

Avoided objects

Freudian abstracts Dust and fibres from Freud's couch



possible consequences for life in those systems. Alternatively, if the enantiomeric excess somehow arose during chemical evolution within Murchison itself, no correlation between meteorites is demanded.

Although these enantiomeric excesses suggest a link between extraterrestrial organic molecules and the origin of terrestrial life, the connection is by no means certain. A preference may have arisen during prebiotic or biotic evolution as well. For example, consider peptide nucleic acid (PNA), a candidate DNA precursor molecule. It has been shown that an otherwise achiral PNA strand can have its chirality fixed by the presence of an L- or D-lysine residue attached to its end¹³. One can imagine a picture in which this random 'seeding' of chirality led to a chiral preference in either prebiotic chemistry or early life¹⁴. So even if the Solar System was born with a preference for L-amino acids, there may have been other opportunities for chiral choices to be made, and our understanding is far too limited to know whether these subsequent choices might have dominated any initial enantiomeric bias.

Finally, Lederberg¹⁵ suggested in 1965 that one criterion for detecting extraterres-

trial life in the Solar System might be to search for enantiomeric excesses. Mounting evidence for (evidently non-biological) enantiomeric excesses in the Murchison meteorite means that this criterion alone may be less useful than we had hoped — whether within martian meteorites, on the martian surface, or elsewhere. □

Christopher F. Chyba is in the Department of Planetary Sciences, University of Arizona, Tucson, Arizona 85721, USA.

- Engel, M. H. & Macko, S. A. *Nature* **389**, 265–268 (1997).
- Kvenvolden, K. *et al.* *Nature* **228**, 923–926 (1970).
- Kvenvolden, K. A., Lawless, J. G. & Ponamperuma, C. *Proc. Natl Acad. Sci. USA* **68**, 486–490 (1971).
- Engel, M. H. & Nagy, B. *Nature* **296**, 837–840 (1982).
- Bada, J. L. *et al.* *Nature* **301**, 494–497 (1983).
- Pillinger, C. T. *Nature* **296**, 802 (1982).
- Engel, M. H., Macko, S. A. & Silfer, J. A. *Nature* **348**, 47–49 (1990).
- Cronin, J. R. & Chang, S. in *The Chemistry of Life's Origins* (eds Greenberg, J. *et al.*) 209–258 (Kluwer, Amsterdam, 1993).
- Zhao, M. & Bada, J. L. *Nature* **339**, 463–465 (1989).
- Cronin, J. R. & Pizzarello, S. *Science* **275**, 951–965 (1997).
- Bonner, W. A. *Origins of Life* **21**, 59–111 (1991).
- Chyba, C. F. & Sagan, C. in *Comets and the Origin and Evolution of Life* (eds Thomas, P. J., Chyba, C. F. & McKay, C. P.) 147–173 (Springer, New York, 1997).
- Wittung, P. *et al.* *Nature* **368**, 561–563 (1994).
- Cohen, J. *Science* **267**, 1265–1266 (1995).
- Lederberg, J. *Nature* **207**, 9–13 (1965).

Binocular vision

Seeing in reverse

Ian P. Howard

It is remarkable that we perceive a three-dimensional world using two-dimensional images. Artists discovered that depth perception arises from perspective, shading and occlusion of distant objects by closer ones. Yet it was not until Wheatstone invented the stereoscope, in 1838, that it was realized that differences in the positions of images in our two eyes — binocular disparities — also create impressions of depth¹. (One of Wheatstone's stereograms is shown in Fig. 1.)

Recently, people have become fascinated by autostereograms, in which a three-dimensional scene emerges when the eyes misconverge on an apparently random array of dots. We now have stereoscopic virtual reality and stereoscopic 'vision' in robots. But what are the brain mechanisms that are involved in stereopsis (binocular vision)²? An advance in our understanding of these mechanisms comes from results by Cumming and Parker³ and by Masson *et al.*⁴, on pages 280 and 283, respectively, of this issue.

When we look at a small object, the images of it fall on corresponding points in the two retinas. Neural signals from these corresponding points then converge on the same 'binocular cell' in the primary visual cortex (V1) of the brain⁵. As a result, the images fuse and we see one object. But if the images differ widely in shape, orientation or luminance contrast, they rival one another

rather than fuse. When similar images fall on slightly different retinal regions, we see one object standing out in depth (Fig. 1). This is the magic of stereoscopic vision. The simplest idea is that optic-nerve fibres serving slightly different regions (receptive fields) in each of the two retinas converge onto the same binocular cell. Such a cell is a 'disparity detector', because it responds best when the images in the two eyes have a disparity that matches the difference in the positions of the two receptive fields⁶. Different disparity detectors respond to different signs and magnitudes of disparity. Zero-disparity detectors respond to objects in the plane of fixation, crossed-disparity detectors react to nearer objects, and uncrossed-disparity

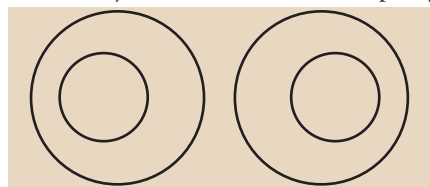


Figure 1 One of the original stereograms constructed by Wheatstone in 1838. When the displays are fused by convergence, the inner ring appears nearer than the outer ring, because its images have a crossed disparity. When the displays are fused by divergence, the inner ring appears more distant, because its images have an uncrossed disparity.

detectors respond to more distant objects. This is a position-disparity mechanism. Disparities may also be detected by differences in the distribution of excitatory and inhibitory regions within two receptive fields that do not differ in position. This is a phase-disparity mechanism.

A simple disparity detector that receives direct inputs from the two eyes does not produce a pure disparity signal — one that is not influenced by incidental changes in the stimulus. For example, the signal for black bars is not the same as that for white bars, and simple detectors are sensitive to slight changes in the location of the object⁷. In the cat, more robust disparity signals are produced by so-called complex cells at a later stage of processing. But, even for some complex detectors, the sign of the disparity signal is inverted when the contrast sign is reversed such that white regions in one eye are superimposed on black regions in the other⁸.

Cumming and Parker³ now report that most binocular cells in V1 of the monkey produce a normal disparity signal to images with the same sign of contrast, but an inverted signal to reversed-contrast images. Visually, images with reversed contrast evoke rivalry, but they do not create an impression of depth (see Fig. 1 on page 281). These results are important because they show that disparity detectors in V1 of the primate respond to a type of disparity that is not used directly for depth perception. Furthermore, they indicate that (contrary to the generally held view), cells in V1 cannot detect when images are in register — this function presumably occurs in a higher visual centre.

What are these inverted-disparity signals used for? Two or more simple detectors could feed into a complex detector which accepts only same-contrast signals. Other complex cells could use reversed-contrast signals to evoke rivalry and help to indicate whether complex images in the two eyes are in proper register (that is, whether they have a maximum proportion of same-contrast edges). But periodic stimuli, even when in register, can contain regions of opposite contrast. Under these circumstances, complex cells that accept opposite-contrast stimuli could use them to improve their response⁸.

When we look at a nearby object our eyes converge, and when we look at a far object our eyes diverge. These eye movements bring the images of the object of interest, as far as possible, onto corresponding points, and simplify the task of detecting disparities due to relative depth. Given that reversed-contrast images produce inverted disparity signals, do they also evoke reversed vergence?

To answer this question, Masson *et al.*⁴ measured the vergence responses of monkeys and humans to arrays of overlapping reversed-contrast disks. Vergence occurred in the opposite direction to that evoked by same-

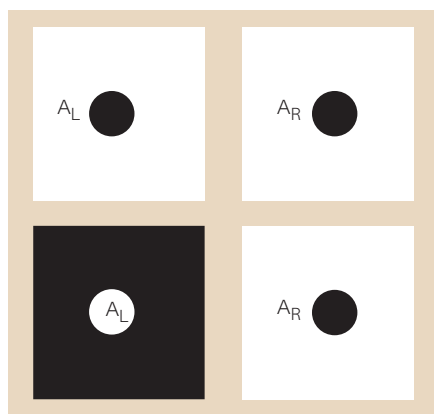


Figure 2 Binocular combination of the two upper displays induces convergence because of the crossed disparity of the black disks. In the two lower displays, there are matching edges (A_L and A_R) with an uncrossed disparity which could induce divergence. Cumming and Parker³ and Masson *et al.*⁴ argue that edge-matching is not the crucial factor that determines the response of the visual system to arrays of reversed-contrast disks.

contrast images. Both normal and reversed vergence occurred within 60–80 milliseconds after the stimulus was presented. Responses this rapid must involve low-level neural processing, because higher levels of processing involve the transmission of signals through several time-consuming stages. The reversed-contrast display produced no impression of depth. These results fit nicely with Cumming and Parker's finding that reversed-contrast images produce inverted-disparity signals in V1 which are not used directly in depth perception. Masson *et al.* add the important finding that depth perception and short-latency vergence can be dissociated.

Masson *et al.* used overlapping 2°-wide disks, so could the reversed vergence have been evoked by disparity between matching edges (of the type shown in Fig. 2), rather than by reversal of contrast phase in the receptive fields of disparity detectors? One would have to assume that closer matching edges have greater control over vergence than more distant matching edges, or non-matching edges. Moreover, predictions are difficult with random multiple-disk displays. The authors of both papers provide theoretical reasons why edge-matching does not explain their data. Masson *et al.* found



Figure 3 One could investigate the response of disparity detectors and of the vergence system to single edges with reversed contrast by presenting a long black bar to one eye and a long white bar to the other eye, with various initial disparities between the two oppositely signed edges.

that small, low-density reversed-contrast disks evoke vergence in the same direction as matching disks, and that these displays produce an impression of depth. These vergence responses occurred with a delay of 110 milliseconds, which suggests the involvement of a higher-level neural mechanism. Images that differ in shape or orientation can also evoke transient vergence in the normal direction⁹. Cumming and Parker still obtained inverted-disparity signals when disparity was larger than the diameter of the stimulus elements. Further work is needed to resolve this apparent contradiction. But the stimulus shown in Fig. 3 — in which there are only contrast edges of one sign in each eye

— might provide a way to study responses to non-matching edges in isolation. □

Ian P. Howard is at the Centre for Vision Research, York University, North York, Ontario M3J 1P3, Canada.

1. Wheatstone, C. *Phil. Trans. R. Soc. Lond.* **128**, 371–394 (1838).
2. Howard, I. P. & Rogers, B. J. *Binocular Vision and Stereopsis* (Oxford Univ. Press, New York, 1995).
3. Cumming, B. G. & Parker, A. J. *Nature* **389**, 280–283 (1997).
4. Masson, G. S., Busetini, C. & Miles, F. A. *Nature* **389**, 283–286 (1997).
5. Hubel, D. H. & Wiesel, T. N. *J. Physiol. Lond.* **160**, 106–154 (1962).
6. Barlow, H. B., Blakemore, C. & Pettigrew, J. D. *J. Physiol. Lond.* **193**, 327–342 (1967).
7. Qian, N. *Vision Res.* **37**, 1811–1827 (1997).
8. Ohzawa, I., DeAngelis, G. C. & Freeman, R. D. *J. Neurophysiol.* **77**, 2879–2910 (1997).
9. Jones, R. *Ann. J. Optom. Physiol. Opt.* **57**, 640–644 (1980).

Apoptosis

Clues in the p53 murder mystery

Andrew Wyllie

The paper by Polyak *et al.* on page 300 of this issue¹ is a bold step towards understanding why expression of the oncosuppressor protein p53 leads to cell death. The issue is important because p53 forms part of a 'fast-track' connection between nuclear DNA damage and the complex but now relatively well-defined cellular apparatus that is responsible for death by apoptosis (Fig. 1, overleaf).

Understanding the elements in this connection would give insight into the mode of action of ionizing radiation and many drugs used in cancer therapy. Further, there is some — albeit incomplete — evidence that, through this function, p53 commits to death those cells that have sustained DNA damage from mutagens, so purging tissues of the founder cells of cancer^{2,3}. p53 is also apparently responsible for cell deletion following hypoxia⁴. Mechanistically, this might be merely a restatement of the theme of DNA damage, because hypoxia, particularly if it is of fluctuating extent, can lead to bursts of production of reactive oxygen species (ROS), which in turn can engender DNA strand breaks. Conceptually, however, it raises the possibility that pharmacological regulation of p53 action might significantly modify the extent of hypoxic tissue injury in myocardial infarction and stroke⁵.

Death is not the only possible outcome of p53 expression after a cell has been injured. In many circumstances, defined in part by cell lineage but also by other, ill-understood factors, the injured cells survive, but are retained in the G1 phase of the cell cycle, sometimes for long periods⁶. The basis for this arrest is induction by p53 of the cyclin kinase inhibitor p21^{waf-1}. Results from p21 knockout animals and other experiments, however, show unequivocally that p21 is not required for post-injury, p53-dependent apoptosis^{7,8}. This implies that, downstream of p53 activa-

tion, there is a means of making the ultimate choice between life and death, but the molecular basis of this choice is enigmatic⁹.

Part of the excitement aroused by the work of Polyak *et al.* is the power of their methodology to find new p53-dependent transcripts that might be cell-specific. A human colorectal cancer cell line, DLD-1, was chosen because it has no functional p53 itself (in common with many such lines), but enters apoptosis within a day or so of contrived expression of wild-type p53, introduced in an adenovirus vector. The transcripts in the parental DLD-1 cells and the p53-expressing, virally infected derivatives were then compared in an unusually comprehensive manner using SAGE (serial analysis of gene expression)¹⁰. Of more than 7,000 transcripts interrogated, only 34 differed in expression between the two cell populations by more than ten times, 14 with greater and 20 with lesser expression in the p53-expressing cells. A series of experiments with one of the p53-induced genes confirmed that its promoter contains an effective p53-binding site, accessible for transcriptional regulation by wild-type p53.

So what are the new p53-induced genes, and what do they tell us about p53-dependent apoptosis? Amongst the 14 was one gene already known to be p53-dependent after injury — p21^{waf-1}. Polyak *et al.* consider that several of the others might influence the availability of intracellular ROS, either through synthesis of scavengers such as glutathione, or as homologues of redox enzyme systems. This is a seductive idea, for ROS are powerful killers, capable of setting up destructive molecular chain reactions in both lipids and nucleic acids, and they are used in precisely this way for the killing of bacteria by cells of the myeloid lineage. Moreover, there is striking new evidence¹¹ for the importance of signals from mitochondria in the initiation of

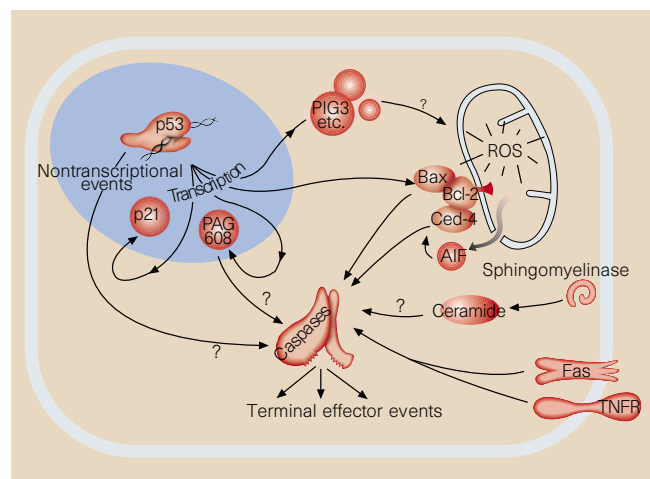


Figure 1 Summary of some of the current thinking on the mechanisms of apoptosis. Stimuli arising within the nucleus (top left) are distinguished from those related to the mitochondrion (top right) and the cell membrane (bottom right). Ultimately, the stimuli converge on the process of activation of caspase enzymes, whose many substrates are thought to account for the terminal events of apoptosis.

apoptosis; these organelles are well placed to be sensors of oxidation damage.

Thus, in at least some circumstances, apoptosis is preceded by precipitous collapse of mitochondrial membrane potential, through loss of selective ion permeability (the so-called permeability transition); mitochondria of cells induced to enter apoptosis release apoptosis-initiating factors (AIFs or Apafs), which trigger the latent activity of the caspases — cysteine proteases whose action drives most if not all of the terminal effector events of apoptosis; Bcl-2, one of the major endogenous anti-apoptotic molecules, is anchored to the outer leaflet of the mitochondrial membrane, where it may regulate the transmembrane flow of ions and other molecules; and two powerful pro-apoptotic molecules, Bax and the nematode Ced-4 protein, dock in quiescent form onto mitochondrial Bcl-2, but are active on release¹².

So maybe the p53-induced, death-specifying transcripts generate a burst of ROS on the mitochondrial membrane, permitting them to both release the safety catch (Bcl-2) and pull the triggers (Bax, Ced-4, AIF) of the terminal effector pathway of apoptosis (Fig. 1). Polyak *et al.* present experiments that show that the time course of events in doomed DLD-1 cells is compatible with this scheme, as is the effect of a variety of inhibitors of transcription, ROS generation and the mitochondrial permeability transition.

Yet the critical death-determining events may still prove elusive. Polyak *et al.* observe that all of the 14 p53-induced genes they have identified are also transactivated in cancer cell lines which, unlike DLD-1, undergo growth arrest rather than apoptosis on expression of p53. There is therefore no guarantee that any of these genes is death-specific. It is true that inhibitors of ROS block, or at least delay, p53-induced death in DLD-1 cells, but such inhibitor experiments have sometimes proved unreliable witnesses.

A problem also arises from the failure to achieve apoptosis after engineered expression of one of the new genes, *PIG3* (for p53-induced gene). *PIG3* protein is a close homo-

logue of a plant oxidoreductase enzyme, and of the 14 p53-induced genes it is arguably the most directly implicated in possible ROS production. Moreover, there is powerful evidence that Bcl-2 rescues doomed cells very effectively even in the complete absence of ROS^{13,14}, and evidence too that the apoptosis of at least some cell types may not involve p53 transactivation at all¹⁵. Finally, there is the question of whether short-term expression of p53 in a cell line such as DLD-1, in which p53 function has been absent for many generations, will necessarily give a complete picture of the genes that are responsive to p53 in more physiological circumstances.

So the long-sought agents that select apoptosis in response to p53 may have indeed been found, singly or in combination, amongst the 14 p53-induced genes identified by Polyak *et al.* Or they may be amongst the 20 new genes still to be worked on that are downregulated by p53. Or they may include another recently identified candidate, the nucleolar-localizing zinc-finger protein PAG608, which unlike *PIG3*, does induce apoptosis when transiently overexpressed¹⁶. Perhaps, however, the game of molecular whodunit is not yet played out. Are all p53-dependent deaths effected by the same means? And does the murder take place in the nucleus or the mitochondrion? □

Andrew Wyllie is in the Sir Alastair Currie CRC Laboratories, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.

1. Polyak, K., Xia, Y., Zweier, J. L., Kinzler, K. W. & Vogelstein, B. *Nature* **389**, 300–305 (1997).
2. Griffiths, S. D. *et al.* *Oncogene* **14**, 523–532 (1996).
3. Clarke, A. R., Howard, L. A., Harrison, D. J. & Winton, D. J. *Oncogene* **14**, 2015–2018 (1997).
4. Graeber, T. G. *et al.* *Nature* **379**, 88–91 (1996).
5. Hara, H. *et al.* *Proc. Natl Acad. Sci. USA* **94**, 2007–2012 (1997).
6. Linke, S. P. *et al.* *Cancer Res.* **57**, 1171–1179 (1997).
7. Deng, C. X. *et al.* *Cell* **82**, 675–684 (1995).
8. Attardi, L. D. *et al.* *EMBO J.* **15**, 3693–3701 (1996).
9. Polyak, K. *et al.* *Genes Dev.* **10**, 1945–1952 (1996).
10. Zhang, L. *et al.* *Science* **276**, 1268–1272 (1997).
11. Kroemer, G. *Nature Med.* **3**, 614–620 (1997).
12. Zou, H. *et al.* *Cell* **90**, 405–413 (1997).
13. Shimizu, S. *et al.* *Nature* **374**, 811–813 (1995).
14. Jacobson, M. D. & Raff, M. C. *Nature* **374**, 814–816 (1995).
15. Caelles, C., Helmberg, A. & Karin, M. *Nature* **370**, 220–223 (1994).
16. Israeli, D. *et al.* *EMBO J.* **16**, 4384–4392 (1997).

Daedalus

Back to the womb

Motherhood, particularly in the early stages, is often a draining ordeal. All new babies have to be fed and cleaned frequently throughout the day and night. But while some are fairly placid and accepting, others are howling terrors. Daedalus feels that technology can help.

The traditional way of soothing a baby, he notes, is to rock its cradle. This recreates a feature of the womb — its steady oscillation at about 1 Hz from the mother's heartbeat. Indeed, heartbeat sounds played to babies also have a soothing effect. So Daedalus is taking the idea to extremes. He wants to recreate the entire vibrational and acoustic environment of an unborn baby, from less than 1 Hz up to perhaps 40 kHz — young ears have a very wide frequency response. It would include internal body noises such as peristalsis, the mother's voice (which also has a uterine familiarity) and domestic noises from outside.

DREADCO technicians are adapting IUD contraceptive technology for the job. They are fitting an IUD with miniature wideband microphones, motion sensors and a radio transmitter. It could be inserted for recording either before the pregnancy was undertaken, or soon after the baby had been born. The resulting tape will be used to drive DREADCO's 'wideband cradle'. This will wrap the baby completely in its old familiar uterine vibrational and acoustic environment.

With luck, the most fractious infant will be soothed by this nostalgic virtual reality. But if not, Daedalus has another trick up his sleeve. In the womb, the baby's nose is filled with amniotic fluid. It must smell its liquid surroundings; and smell is the most immediate and emotional of all the senses. Birth withdraws this calming aroma for good. So DREADCO chemists are analysing amniotic fluid, and studying its variation between mothers, to replicate it in aerosol form. Sprayed around the cradle, it will powerfully reinforce the soothing womb nostalgia.

There may be adult applications too. Some psychologists believe that the trauma of birth can do lasting subconscious damage, only healed by recalling its repressed memory. To this end, 're-birthing' ceremonies are sometimes practised. Uterine recall therapy should be greatly aided by original recordings and aerosols from the patient's womb environment. Sadly, this bold idea cannot be checked until the first crop of uterine-recorded babies have grown up and developed psychological problems.

David Jones

Gene therapy – promises, problems and prospects

Inder M. Verma and Nikunj Somia

In principle, gene therapy is simple: putting corrective genetic material into cells alleviates the symptoms of disease. In practice, considerable obstacles have emerged. But, thanks to better delivery systems, there is hope that the technique will succeed.

In 1990, the first clinical trials for gene-therapy approaches to combat disease were carried out. Conceptually, the technique involves identifying appropriate DNA sequences and cell types, then developing suitable ways in which to get enough of the DNA into these cells. With efficient delivery, the therapeutic prospects range from tackling genetic diseases and slowing the progression of tumours, to fighting viral infections and stopping neurodegenerative diseases. But the problems — such as the lack of efficient delivery systems, lack of sustained expression, and host immune reactions — remain formidable challenges.

Although more than 200 clinical trials are currently underway worldwide, with hundreds of patients enrolled, there is still no single outcome that we can point to as a success story. To explore why this is the case, we will use our own experience and other examples to look at the many technical, logistical and, in some cases, conceptual hurdles that need to be overcome before gene therapy becomes routine practice in medicine.

At present, gene therapy is being contemplated only on somatic (essentially, non-reproductive) cells. Although many somatic tissues can receive therapeutic DNA, the choice of cell usually depends on the nature of the disease. Sometimes a clear definition of the target cell is needed. For example, the gene that is defective in cystic fibrosis has been identified, and clinical trials to deliver DNA as an aerosol into the lung have already begun¹. Although cystic fibrosis is manifest in this organ, it is still not clear that delivery of a correcting gene by this method will reach the right type of cell. On the other hand, to correct blood-clotting disorders such as haemophilia, all that is needed is a therapeutic level of clotting protein in the plasma². This protein may be supplied by muscle or liver cells, fibroblasts, or even blood cells^{3–5}. The choice of tissue in which to express the therapeutic protein will also ultimately depend on considerations such as the efficiency of gene delivery, protein modifications, immunological

status, accessibility and economics.

We also need to consider how much of the therapeutic protein should be delivered. In haemophilia B, which is caused by a deficiency of a blood-clotting protein called factor IX, giving patients just 5% of the normal circulating levels of this protein can substantially improve their quality of life². Most people have about 5 µg of factor IX per millilitre of plasma, produced by the 10^{13} cells that make up the liver. So we need to deliver a correcting gene to 5×10^{11} cells — that is, 5% of liver cells. Alternatively, fewer liver cells would need to be modified if more factor IX could be produced per cell, without being deleterious. In the brain, however, gene transfer to just a few hundred cells

could considerably benefit patients with neurological disease. And finally, we can consider the transfer of genes to a handful of stem (or progenitor) cells, which grow and divide to generate millions of progeny. The range in the number of cells that this technology has to cover is vast.

The Achilles heel of gene therapy is gene delivery, and this is the aspect that we will concentrate on here. Thus far, the problem has been an inability to deliver genes efficiently and to obtain sustained expression. There are two categories of delivery vehicle ('vector'). The first comprises the non-viral vectors, ranging from direct injection of DNA to mixing the DNA with polylysine or cationic lipids that allow the gene to cross the cell membrane. Most of these approaches suffer from poor efficiency of delivery and transient expression of the gene⁶. Although there are reagents that increase the efficiency of delivery, transient expression of the transgene is a conceptual hurdle that needs to be addressed.

Most of the current gene-therapy approaches make use of the second category — viral vectors. Importantly, the viruses used have all been disabled of any pathogenic effects. The use of viruses is a powerful technique, because many of them have evolved a specific machinery to deliver DNA to cells. However, humans have an immune system to fight off the virus, and our attempts to deliver genes in viral vectors have been confronted by these host responses. ▶

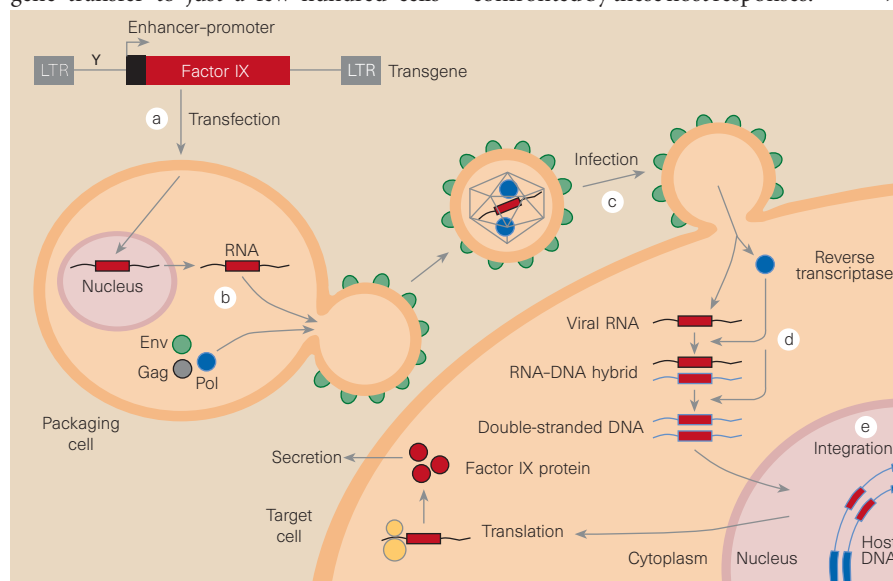


Figure 1 To create the retroviral vectors that are used in gene therapy, the life-cycles of their naturally occurring counterparts are exploited. **a**, The transgene (in this case, the gene for factor IX) in a vector backbone is put into a packaging cell, which expresses the genes that are required for viral integration (*gag*, *pol* and *env*). **b**, The transgene is incorporated into the nucleus, where it is transcribed to make vector RNA. This is then packaged into the retroviral vector, which is shed from the packaging cell. **c**, The vector is delivered to the target cell by infection. The membrane of the viral vector fuses with the target cell, allowing the vector RNA to enter. **d**, The virally encoded enzyme reverse transcriptase converts the vector RNA into an RNA–DNA hybrid, and then into double-stranded DNA. **e**, The vector DNA is integrated into the host genome, then the host-cell machinery will transcribe and translate it to make RNA and, in this case, factor IX protein.

LTR, long terminal repeat; ψ , packaging sequence.

Retroviral vectors

Retroviruses are a group of viruses whose RNA genome is converted to DNA in the infected cell. The genome comprises three genes termed *gag*, *pol* and *env*, which are flanked by elements called long terminal repeats (LTRs). These are required for integration into the host genome, and they define the beginning and end of the viral genome. The LTRs also serve as enhancer–promoter sequences — that is, they control expression of the viral genes. The final element of the genome, called the packaging sequence (ψ), allows the viral RNA to be distinguished from other RNAs in the cell (Fig. 1)⁷.

By manipulating the viral genome, viral genes can be replaced with transgenes — such as the gene for factor IX (Table 1). Transcription of the transgene may be under the control of viral LTRs or, alternatively, enhancer–promoter elements can be engineered in with the transgene. The chimaeric genome is then introduced into a packaging cell, which produces all of the viral proteins (such as the products of the *gag*, *pol* and *env* genes), but these have been separated from the LTRs and the packaging sequence. So, only the chimaeric viral genomes are assembled to generate a retroviral vector. The culture medium in which these packaging cells have been grown is then applied to the target cells, resulting in transfer of the transgene. Typically, a million target cells on a culture dish can be infected with one millilitre of the viral soup.

A critical limitation of retroviral vectors is their inability to infect non-dividing cells⁸, such as those that make up muscle, brain, lung and liver tissue. So, when possible, the cells from the target tissue are removed,

grown *in vitro*, and infected with the recombinant retroviral vector. The target cells producing the foreign protein are then transplanted back into the animal. This procedure has been termed ‘*ex vivo* gene therapy’ and our group has used it to infect mouse primary fibroblasts or myoblasts (connective-tissue and muscle precursors, respectively) with retroviral vectors producing the factor IX protein. But within five to seven days of transplanting the infected cells back into mice, expression of factor IX is shut off^{3,5,9}. This transcriptional shut-off has even been observed in mice lacking a functional immune system (nude mice), and it cannot be due to cell loss or gene deletion⁵ because the transplanted cells can be recovered.

What is the mechanism of this unexpected but intriguing problem? We do not yet know, but the exceptions may provide some clues. To obtain sustained expression in mouse muscle following the transplantation of infected myoblasts, we used the muscle creatine kinase enhancer–promoter to control transcription of the transgene. Unfortunately, this is a weak promoter, and only low levels of protein were produced. So, we generated a chimaeric vector in which the muscle creatine kinase enhancer was linked to a strong promoter from cytomegalovirus. Using this vector, sustained and high levels of factor IX were expressed when the infected myoblasts were transplanted back into mice. Remarkably, these expression levels remained unchanged for more than two years (the life of the animal). So we can override the ‘off switch’ and achieve higher levels of expression by using an appropriate enhancer–promoter combination. But the search for such combinations is a case

of trial and error for a given type of cell.

Another formidable challenge to the *ex vivo* approach is the efficiency of transplantation of the infected cells. Attempts to repeat the long-term myoblast transplantation in haemophilic dogs led to only short-term expression, because the infected dog myoblasts could not fuse with the muscle fibres. So perhaps successful animal models will prove inadequate when the same protocols are extended to humans. Moreover, these models are based on inbred animals — the outbred human population, with individual variation, will add yet another degree of complexity. The haematopoietic (blood-producing) system may offer an advantage for *ex vivo* gene therapy because resting stem cells can be stimulated to divide *in vitro* using growth factors and the transplantation technology is well established. But there is still a lack of good enhancer–promoter combinations that allow sustained production of high levels of protein in these cells.

Another problem is the possibility of random integration of vector DNA into the host chromosome. This could lead to activation of oncogenes or inactivation of tumour-suppressor genes. Although the theoretical probability of such an event is quite low, it is of some concern (see section on clinical trials).

Lentiviral vectors

Lentiviruses also belong to the retrovirus family, but they can infect both dividing and non-dividing cells¹⁰. The best-known lentivirus is the human immunodeficiency virus (HIV), which has been disabled and developed as a vector for *in vivo* gene delivery. Like the simple retroviruses, HIV has the three *gag*, *pol* and *env* genes, but it also carries genes for six accessory proteins termed *tat*, *rev*, *vpr*, *vpu*, *nef* and *vif*¹¹.

Using the retrovirus vectors as a model, lentivirus vectors have been made, with the transgene enclosed between the LTRs and a packaging sequence¹². Some of the accessory proteins can be eliminated without affecting production of the vector or efficiency of infection. The *env* gene product would restrict HIV-based vectors to infecting only cells that express a protein called CD4¹ so, in the vectors, this gene is substituted with *env* sequences from other RNA viruses that have a broader infection spectrum (such as glycoprotein from the vesicular stomatitis virus). These vectors can be produced — albeit on a small scale at the moment — at concentrations of >10⁷ virus particles per ml (ref.12).

When lentivirus vectors are injected into rodent brain, liver, muscle, eye or pancreatic-islet cells, they give sustained expression for over six months — the longest time tested so far^{13,14}. Unlike the prototypical retroviral vectors, the expression is not subject to ‘shut off’. Little is known about the possible immune problems associated with lentiviral vectors, but injection of 10⁷ infectious units

Table 1 **Candidate diseases for gene therapy**

Disease	Defect	Incidence	Target cells
Genetic			
Severe combined immunodeficiency (SCID/ADA)	Adenosine deaminase (ADA) in ~25% of SCID patients	Rare	Bone-marrow cells or T lymphocytes
Haemophilia	A Factor VIII deficiency	1:10,000 males	Liver, muscle, fibroblasts or bone-marrow cells
	B Factor IX deficiency	1:30,000 males	
Familial hypercholesterolaemia	Deficiency of low-density lipoprotein (LDL) receptor	1:1 million	Liver
Cystic fibrosis	Faulty transport of salt in lung epithelium. Loss of <i>CFTR</i> gene	1:3,000 Caucasians	Airways in the lungs
Haemoglobinopathies: thalassaemias/sickle-cell anaemia	Structural defects in α - or β -globin gene	1:600 in certain ethnic groups	Bone-marrow cells, giving rise to red blood cells
Gaucher's disease	Defect in the enzyme glucocerebrosidase	1:450 in Ashkenazi Jews	Bone-marrow cells, macrophages
α_1 -antitrypsin deficiency: inherited emphysema	Lack of α_1 -antitrypsin	1:3,500	Lung or liver cells
Acquired			
Cancer	Many causes, including genetic and environmental	1 million/year in USA	Variety of cancer-cell types; liver, brain, pancreas, breast, kidney
Neurological diseases	Parkinson's, Alzheimer's, spinal-cord injury	1 million Parkinson's and 4 million Alzheimer's patients in USA	Direct injection in the brain, neurons, glial cells, Schwann cells
Cardiovascular	Restinosis, arteriosclerosis	13 million in USA	Arteries, vascular endothelial cells
Infectious diseases	AIDS, hepatitis B	Increasing numbers	T cells, liver, macrophages

does not elicit the cellular immune response at the site of injection. Furthermore, there seems to be no potent antibody response. So, at present, lentiviral vectors seem to offer an excellent opportunity for *in vivo* gene delivery with sustained expression.

Adenoviral vectors

The adenoviruses are a family of DNA viruses that can infect both dividing and non-dividing cells, causing benign respiratory-tract infections in humans¹¹. Their genomes contain over a dozen genes, and they do not usually integrate into the host DNA. Instead, they are replicated as episomal (extrachromosomal) elements in the nucleus of the host cell. Replication-deficient adenovirus vectors can be generated by replacing the *E1* gene — which is essential for viral replication — with the gene of interest (for example, that for factor IX) and an enhancer–promoter element. The recombinant vectors are then replicated in cells that express the products of the *E1* gene, and they can be generated in very high concentrations ($>10^{11}$ – 10^{12} adenovirus particles per ml)¹⁵.

Cells infected with recombinant adenovirus can express the therapeutic gene but, because essential genes for viral replication are deleted, the vector should not replicate. These vectors can infect cells *in vivo*, causing them to express very high levels of the transgene. Unfortunately, this expression lasts for only a short time (5–10 days post-infection). In contrast to the retroviral vectors, long-term expression can be achieved if the recombinant adenoviral vectors are introduced into nude mice, or into mice that are given both the adenoviral vector and immunosuppressing agents¹⁶. This indicates that the immune system is behind the short-term expression that is usually obtained from adenoviral vectors.

The immune reaction is potent, eliciting both the cell-killing ‘cellular’ response and the antibody-producing ‘humoral’ response. In the cellular response, virally infected cells are killed by cytotoxic T lymphocytes^{16,17}. The humoral response results in the generation of antibodies to adenoviral proteins, and it will prevent any subsequent infection if the animal is given a second injection of the recombinant adenovirus. Unfortunately for gene therapy, most of the human population will probably have antibodies to adenovirus from previous infection with the naturally occurring virus.

It is possible that the target cell contains factors that might trigger the synthesis of adenoviral proteins, leading to an immune response. To try to get around this problem, second-generation adenoviral vectors were developed, in which additional genes that are implicated in viral replication were deleted. These vectors showed longer-term expression, but a decline after 20–40 days was still apparent¹⁸. This idea has now been taken fur-

What makes an ideal vector?

All of the current methods of gene delivery — whether viral or non-viral — have some limitation. So, the choice of vector will often be dictated by the need. If expression of the gene is required for only a short time (for example, expression of a toxic gene-product in cancer cells), then the adenoviral vectors are ideal. But if sustained expression is needed (such as for most genetic diseases), then an integrating vector

without attendant immunological problems is more desirable. An ideal vector may have to borrow properties from both viral and synthetic systems, and it should have:

- High concentration ($>10^8$ viral particles per ml), allowing many cells to be infected;
- Convenience and reproducibility of production;
- Ability to integrate in a site-specific location in the host chromosome, or

to be successfully maintained as a stable episome;

- A transcriptional unit that can respond to manipulation of its regulatory elements;
- Ability to target the desired type of cell;
- No components that elicit an immune response.

Although no such vector is currently available, all of these properties exist, individually, in disparate delivery systems.

ther with the generation of ‘gut-less’ vectors — all of the viral genes are deleted, leaving only the elements that define the beginning and the end of the genome, and the viral packaging sequence. The transgenes carried by these viruses were expressed for 84 days¹⁹.

There are considerable immunological problems to be overcome before adenoviral vectors can be used to deliver genes and produce sustained expression. The incoming adenoviral proteins that package DNA can be transported to the cytoplasm where they are processed and presented on the cell surface, tagging the cell as infected for destruction by cytotoxic T cells. So adenoviral vectors are extremely useful if expression of the transgene is required for short periods of time. One promising approach is to deliver large numbers of adenoviral vectors containing genes for enzymes that can activate cell killing, or immunomodulatory genes, to cancer cells. In this case, the cellular immune response against the viral proteins will augment tumour killing. Finally, although immunosuppressive drugs can extend the duration of expression, our goal should be to manipulate the vector and not the patient.

Adeno-associated viral vectors

A relative newcomer to the field, adeno-associated virus (AAV) is a simple, non-pathogenic, single-stranded DNA virus. Its two genes (*cap* and *rep*) are sandwiched between inverted terminal repeats that define the beginning and the end of the virus, and contain the packaging sequence²⁰. The *cap* gene encodes viral capsid (coat) proteins, and the *rep* gene product is involved in viral replication and integration. AAV needs additional genes to replicate, and these are provided by a helper virus (usually adenovirus or herpes simplex virus).

The virus can infect a variety of cell types, and — in the presence of the *rep* gene product — the viral DNA can integrate preferen-

tially²⁰ into human chromosome 19. To produce an AAV vector, the *rep* and *cap* genes are replaced with a transgene. Up to 10^{11} – 10^{12} viral particles can be produced per ml, but only one in 100–1,000 particles is infectious. Moreover, preparation of the vector is laborious and, due to the toxic nature of the *rep* gene product and some of the adenoviral helper proteins, we currently have no packaging cells in which all of the proteins can be stably provided. Vector preparations must also be carefully separated from any contaminating adenovirus, and AAV vectors can accommodate only 3.5–4.0 kilobases of foreign DNA — this will exclude larger genes. Finally, we need more information about the immunogenicity of the viral proteins, especially given that 80% of the adult population have circulating antibodies to AAV. These considerations notwithstanding, AAV vectors containing human factor IX complementary DNA have been used to infect liver and muscle cells in immunocompetent mice. The mice produced therapeutic amounts of factor IX protein in their blood for over six months^{21,22}, confirming the promise of AAV as an *in vivo* gene-therapy vector.

Other vectors

Among the other viruses being considered and developed, is herpes simplex virus, which infects cells of the nervous system²³. The virus contains more than 80 genes, one of which (*IE3*) can be replaced to create the vector. Around 10^8 – 10^9 viral particles are produced per ml, but the residual proteins are toxic to the target cell. Additional genes can be deleted, allowing more than one transgene to be included. But if essentially all of the viral proteins are deleted (gut-less vectors), only around 10^6 viral particles are produced per ml. And, again, many people have an immunity to components of herpes simplex virus, having already been infected at some time.

Vaccinia-virus-based vectors have also

been explored, largely for generating vaccines²⁴. The Sindbis and Semliki Forest virus is being exploited as a possible cytoplasmic vector²⁵ which does not integrate into the nucleus. Although most of these systems produce the foreign protein only transiently, they add diversity to the available systems of gene transfer (Table 2).

Clinical trials

Over half of the 200 clinical trials that have been launched in the United States involve therapeutic approaches to cancer. Nearly 30 of them involve correction of monogenic diseases (Table 1) such as cystic fibrosis, α_1 -antitrypsin deficiency and severe combined immunodeficiency (SCID). Most of the trials are phase I (safety) studies and, by and large, the existing delivery systems have no major toxicity problems. Moreover, lessons can be learned from previous clinical trials^{26,27}. For example, seven years ago two patients were enrolled in a trial to correct deficiencies in adenosine deaminase (ADA, which leads to severe immunodeficiency). One of the patients improved, and makes 25% of normal ADA in her T cells, five years after the last infusion of infected T cells (although she is still treated with PEG-ADA; bovine adenosine deaminase mixed with polyethylene glycol). But in the other patient, the infected T cells could not produce enough of the deficient enzyme.

The efficacy of gene therapy cannot be evaluated until patients are completely taken off alternative treatments (in the above example, PEG-ADA). In another trial²⁸, weaning a patient away from PEG-ADA reduced the ability of the T cells to respond *in vitro* to a challenge by pathogens. Clearly, in these cases the retroviral vectors were not optimal, and the number of infected blood-progenitor cells was extremely low. However, these trials did help to establish the technology for generating clinical-grade recombinant retroviral particles, the

procedures for infection and transplantation, and the protocols for monitoring patients and analysing data. The disappointing outcome probably lies in the inefficient gene-delivery system. We need a system in which the vector — containing the ADA gene — is efficiently delivered to progenitor cells, leading to sustained expression of high levels of the ADA protein. But, encouragingly, despite being repeatedly injected with highly concentrated recombinant viruses, the patients have suffered no untoward effects to date.

Future prospects

We now need a renewed emphasis on the basic science behind gene therapy — particularly the three intertwined fields of vectors, immunology and cell biology.

All of the available viral vectors arose from understanding the basic biology of the structure and replication of viruses. Clearly, existing vectors need to be streamlined further (see box on page 241), and vectors that target specific types of cell are being developed. The use of antibody fragments, ligands to cell-specific receptors, or chemical modifications to the vector need to be explored systematically. And advances such as the report — published only last week²⁹ — of the three-dimensional structure of the Env protein from mouse leukaemia virus (a commonly used vector), will facilitate the design of targeted vectors. A better understanding of gene transcription will allow us to design regulatory elements that permit promoter activity to be modulated, and development of tissue-specific enhancer-promoter elements should be vigorously pursued. Finally, we need to begin merging some of the qualities of viral vectors with non-viral vectors, to generate a safe and efficient gene-delivery system.

With respect to immunology, viruses still have many secrets to be unravelled. Viral systems that have evolved to escape immune surveillance can be incorporated into viral

vectors. Some of these are being characterized; for example, the adenoviral E3 protein, the herpes simplex ICP47 protein and the cytomegalovirus US11 protein³⁰. Systems from other pathogens may also be borrowed and incorporated into vectors. In some cases, the correcting protein will be sensed as foreign, eliciting an immune response. Animal models should help us to understand this and, where necessary, to develop strategies for tolerance.

Cell biology is involved because, in many cases, the goal of gene therapy is to correct differentiated cells, such as epithelial cells in cystic fibrosis and lymphoid cells in ADA deficiency. However, because these cells are continuously replaced there has to be either continued therapy or an attempt to target the stem cells. We first need to develop further the technologies for identifying and isolating these cells, to understand their regulation, and to target infection to them *in vivo*.

So how far have we come since clinical trials began? The promises are still great, and the problems have been identified (and they are surmountable). But what of the prospects? Our view is that, in the not too distant future, gene therapy will become as routine a practice as heart transplants are today. □

Inder M. Verma and Nikunj Somia are in the Laboratory of Genetics, The Salk Institute for Biological Studies, La Jolla, California 92037, USA.

Table 2 Comparison of properties of various vector systems

Features	Retroviral	Lentiviral	Adenoviral	AAV	Naked/ lipid-DNA
Maximum insert size	7–7.5 kb	7–7.5 kb	~30 kb	3.5–4.0 kb	Unlimited size
Concentrations (viral particles per ml)	>10 ⁸	>10 ⁸	>10 ¹¹	>10 ¹²	No limitation
Route of gene delivery	<i>Ex vivo</i>	<i>Ex/in vivo</i>	<i>Ex/in vivo</i>	<i>Ex/in vivo</i>	<i>Ex/in vivo</i>
Integration	Yes	Yes	No	Yes/No	Very poor
Duration of expression <i>in vivo</i>	Short	Long	Short	Long	Short
Stability	Good	Not tested	Good	Good	Very good
Ease of preparation (scale up)	Pilot scale up, up to 20–50 l	Not known	Easy to scale up	Difficult to purify, difficult to scale up	Easy to scale up
Immunological problems	Few	Few	Extensive	Not known	None
Pre-existing host immunity	Unlikely	Unlikely, except maybe AIDS patients	Yes	Yes	No
Safety problems	Insertional mutagenesis?	Insertional mutagenesis?	Inflammatory response, toxicity	Inflammatory response, toxicity	None

- Crystal, R. G. *Science* **270**, 404–410 (1995).
- Scriver, C. R. S., Beaudet, A. L., Sly, W. S. & Valle, D. V. *The Metabolic Basis of Inherited Disease* (McGraw-Hill, New York, 1989).
- Dai, Y., Roman, M., Naviaux, R. K. & Verma, I. M. *Proc. Natl. Acad. Sci. USA* **89**, 10892–10895 (1992).
- Kay, M. A. *et al. Science* **262**, 117–119 (1993).
- St Louis, D. & Verma, I. M. *Proc. Natl. Acad. Sci. USA* **85**, 3150–3154 (1988).
- Felgner, P. L. *Sci. Am.* **276**, 102–106 (1997).
- Verma, I. M. *Sci. Am.* **263**, 68–72 (1990).
- Miller, D. G., Adam, M. A. & Miller, A. D. *Mol. Cell. Biol.* **10**, 4239–4242 (1990).
- Palmer, T. D., Rosman, G. J., Osborne, W. R. & Miller, A. D. *Proc. Natl. Acad. Sci. USA* **88**, 1330–1334 (1991).
- Lewis, P., Hensel, M. & Emerman, M. *EMBO J.* **11**, 3053–3058 (1992).
- Field, B. N., Knipe, D. M. & Howley, P. M. (eds) *Virology* (Lippincott-Raven, Philadelphia, PA, 1996).
- Naldini, L. *et al. Science* **272**, 263–267 (1996).
- Miyoshi, H., Takahashi, M., Gage, F. H. & Verma, I. M. *Proc. Natl. Acad. Sci. USA* **94**, 10319–10323 (1997).
- Blömer, U. *et al. J. Virol.* **71**, 6641–6649 (1997).
- Yeh, P. & Perricaudet, M. *FASEB J.* **11**, 615–623 (1997).
- Dai, Y. *et al. Proc. Natl. Acad. Sci. USA* **92**, 1401–1405 (1995).
- Yang, Y., Ertl, H. C. & Wilson, J. M. *Immunity* **1**, 433–442 (1994).
- Engelhardt, J. F., Ye, X., Doranz, B. & Wilson, J. M. *Proc. Natl. Acad. Sci. USA* **91**, 6196–6200 (1994).
- Chen, H. H. *et al. Proc. Natl. Acad. Sci. USA* **94**, 1645–1650 (1997).
- Muzyczka, N. *Curr. Top. Microbiol. Immunol.* **158**, 97–127 (1992).
- Snyder, R. O. *et al. Nature Genet.* **16**, 270–276 (1997).
- Herzog, R. W. *et al. Proc. Natl. Acad. Sci. USA* **94**, 5804–5809 (1997).
- Fink, D. J., DeLuca, N. A., Goins, W. F. & Glorioso, J. C. *Annu. Rev. Neurosci.* **19**, 265–287 (1996).
- Moss, B. *Proc. Natl. Acad. Sci. USA* **93**, 11341–11348 (1996).
- Berglund, P. *et al. Biotechnology (NY)* **11**, 916–920 (1993).
- Blaese, R. M. *et al. Science* **270**, 475–480 (1995).
- Bordignon, C. *et al. Science* **270**, 470–475 (1995).
- Kohn, D. B. *et al. Nature Med.* **1**, 1017–1023 (1995).
- Fass, D. *et al. Science* **277**, 1662–1665 (1997).
- Wiertz, E. J. *et al. Cell* **84**, 769–779 (1996).

Superconductivity enhanced by Hg fission

For the successful application of high-temperature copper oxide superconductors, the problem of the ease of motion of magnetic vortices (quantized flux lines) within the material must be solved. The motion results in finite electrical resistance which prevents the desired loss-free conduction of current¹. We demonstrate a solution to this problem by anchoring the vortices with crystallographic defects induced by fission of mercury atoms in a mercury/copper oxide superconductor.

Easy vortex motion, the origin of which is intrinsic to the material, can be counteracted by correlated disorder in the form of columnar defects². Such defects effectively

localize vortices¹ and expand the range of finite critical current density J_c (marked by the irreversibility line¹ in the field-temperature (H - T) diagram)^{2,3}. One possibly technologically relevant way to produce an effective columnar defect structure is by irradiation with energetic (~ 0.2 – 0.8 GeV) protons, which will induce nuclear fission of sufficiently heavy nuclei⁴. The resulting fission fragments could in turn produce columnar damage tracks⁵, as has been shown for ^{209}Bi nuclei⁶. There is no directionality in the fission process, so the resulting tracks will be highly misaligned (splayed)⁶. The action of splay is to force the entanglement of vortices⁷, expected to reduce thermal activation (creep) of vortices out of columnar tracks and to enhance J_c even further.

In any superconductor, the irreversibility line in the H - T space, above which the vortex matter 'liquefies' and is very mobile¹, is ultimately limited by the transition temperature T_c . So the question arising is whether the fission process can be realized in mercury/copper oxides — materials with the highest T_c found to date^{8,9}. The answer will depend on three factors: the experimentally unknown fission cross-section σ

for a ^{200}Hg fission process launched with fast protons, the as-yet unknown threshold of electronic energy loss rate for columnar track formation⁵ in mercury/copper oxides by the fission fragments, and the efficacy of splayed tracks in these cuprates.

We exposed polycrystalline samples of $\text{HgBa}_2\text{CaCu}_2\text{O}_{6+\delta}$ (Hg-1212)⁹ to 0.8-GeV protons (fluence of $1.52 \times 10^{17} \text{ cm}^{-2}$) at Los Alamos National Laboratory (backed by the Electric Power Research Institute and NSF) using the proton beam from Los Alamos Meson Physics facility at the Weapons Neutron Research branch. The result was the first experimental record of a mercury fission process by which we could install splayed tracks in a Hg-1212 superconductor with transition temperature T_c of roughly 120 K. A cross-sectional transmission electron microscope image of Hg-1212 (Fig. 1a) illustrates the splayed columnar nature of the damage, typical of a fission process.

The effect on current conduction is shown in Fig. 1b, c. Before irradiation the persistent current density J in high magnetic fields falls off rapidly with temperature — by three orders of magnitude at $\mu_0 H = 2$ tesla and $T \approx 60$ K. After irradiation, J is enhanced by orders of magnitude in fields of several tesla and the regime of finite J is extended to $T > 100$ K (higher than in yttrium-, bismuth-, or thallium-based materials^{2,3}). Even for the relatively low proton fluence used here, the large shift in the irreversibility temperature $\Delta T_{\text{irr}} \approx 25$ K is maintained in a field of up to 5.5 tesla (Fig. 1c).

The large shift in T_{irr} is possibly a result of the renormalization of splay angles¹ by sizeable superconductive anisotropy of Hg-1212 (refs 8,9). A uniform splay distribution will not work in a less anisotropic material, such as yttrium-barium-copper oxides, where for splay angles greater than 10° the motion of vortices is enhanced¹⁰. Our results imply that the fission process can be extended to higher proton fluences and to mercury/copper oxides with T_c above 130 K, for which a wire and tape fabrication process has recently been developed¹¹.

L. Krusin-Elbaum, D. Lopez

IBM T. J. Watson Research Center,
Yorktown Heights, New York 10598, USA

J. R. Thompson

Department of Physics, University of Tennessee,
Knoxville, Tennessee 37996, USA
and Oak Ridge National Laboratory,
Oak Ridge, Tennessee 37831, USA

R. Wheeler

The Ohio State University,
Columbus, Ohio 43210, USA

J. Ullmann

Los Alamos National Laboratory,
Los Alamos, New Mexico 87545, USA

C. W. Chu, Q. M. Lin

Texas Center for Superconductivity,
Houston University, Houston, Texas 77204, USA

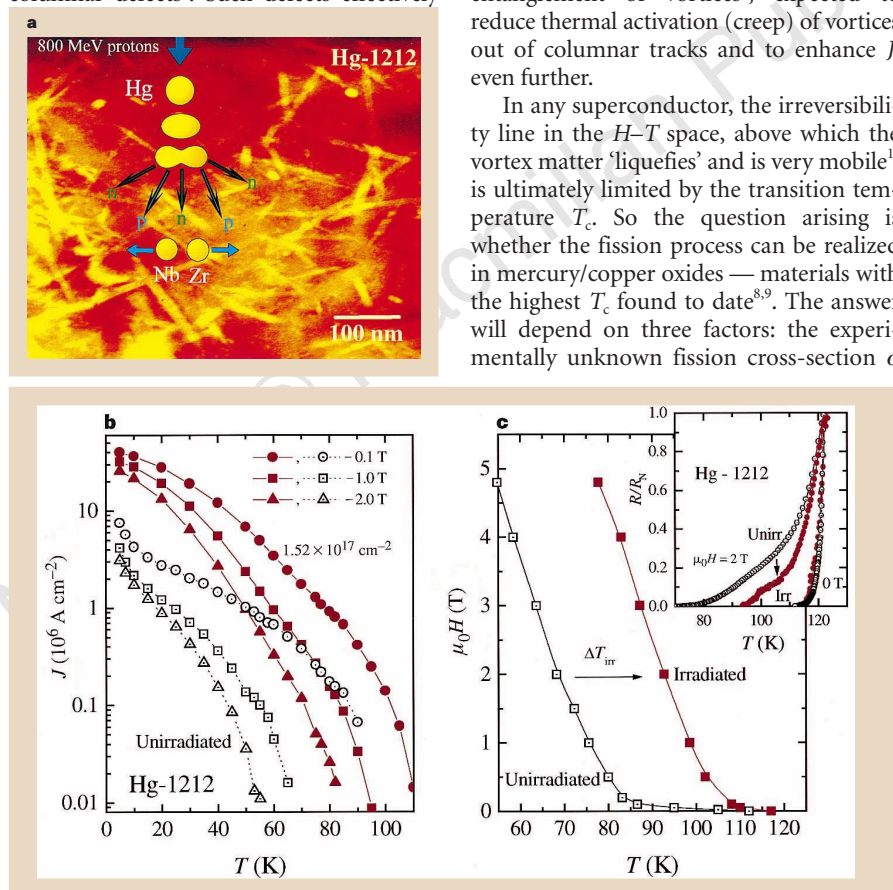


Figure 1 a, Transmission electron micrograph of Hg-1212 bulk superconductor prepared as in ref. 9, irradiated with 0.8-GeV protons. Fission track diameter, ~ 80 – 90 Å; defect density, $\sim 5.9 \times 10^{10} \text{ cm}^{-2}$; $\sigma \sim 110$ mbarns; Hg density, $\sim 5.3 \times 10^{21} \text{ cm}^{-3}$; density of events, 8.9×10^{13} fissions cm^{-3} . Mean track length implied is ~ 6.6 μm , consistent with calculations^{5,6}. Sketch illustrates the fission process. On entering a mercury nucleus, a 0.8-GeV proton collides with nucleons. The nucleus 'boils off' protons and neutrons, which carry off about 0.6 GeV energy. The remaining energy is split between the fission daughters. b, Persistent current density J against temperature for Hg-1212 before and after proton exposure, for three applied magnetic fields. Bulk J is above 10^7 A cm^{-2} . After irradiation J becomes large above 100 K, well above the irreversibility line of unirradiated sample. c, Irreversibility lines of unirradiated and 0.8-GeV proton-irradiated Hg-1212 obtained from the onset of nonlinear current-voltage characteristics. Arrow, ~ 25 K expansion at $\mu_0 H = 2$ T on irradiation. Inset, resistance R (normalized to R at T_c) before and after irradiation. In finite fields onset of linear resistance (dissipation) is shifted to higher temperatures and normalized R at any T is significantly reduced (arrow) by proton irradiation.

- Blatter, G. *et al.* *Rev. Mod. Phys.* **66**, 1125 (1994).
- Civale, L. *et al.* *Phys. Rev. Lett.* **67**, 648–651 (1991).
- Budhani, R. C., Suenaga, M. & Liou, S. H. *Phys. Rev. Lett.* **69**, 3816–3819 (1992).
- Becchetti, F. D. *et al.* *Phys. Rev. C* **28**, 276–279 (1983).
- Toulemonde, M., Bouffard, S. & Studer, F. *Nucl. Instr. Meth. B* **91**, 108–123 (1994).
- Krusin-Elbaum, L. *et al.* *Appl. Phys. Lett.* **64**, 3331–3333 (1994).
- Hwa, T., LeDoussal, P., Nelson, D. R. & Vinokur, V. M. *Phys. Rev. Lett.* **71**, 3545–3548 (1996).
- Schilling, A., Cantoni, M., Guo, J. D. & Ott, H. R. *Nature* **363**, 56–58 (1993).
- Xiong, Q. *et al.* *J. Appl. Phys.* **76**, 7127–7129 (1994).
- Krusin-Elbaum, L. *et al.* *Phys. Rev. Lett.* **76**, 2563–2566 (1996).
- Meng, R. L. *et al.* *Physica C* (in the press).

Glycerol generates turgor in rice blast

Many plant pathogenic fungi are able to penetrate the cuticles of their host plants by elaborating specialized cells known as appressoria^{1,2}. The morphology and development of appressoria have been well studied², but little is known about how these cells are able to breach the tough plant surface. We have now found that the appressoria of rice blast fungus (*Magnaporthe grisea*) use glycerol to generate pressure which ruptures plant cuticles.

Rice blast is the most serious disease of cultivated rice, the staple food for one-third of the world's population^{3,4}. The fungus forms heavily melanin-pigmented appressoria that generate enormous turgor. The pressure is applied as a physical force to break the rice leaf cuticle^{5,6}. Turgor can be as great as 8.0 MPa (modal value, 6.0 MPa)⁵, equivalent to 40 times the pressure in a car tyre and is, to our knowledge, the highest recorded in any living organism⁵.

We extracted the contents of appressoria and biochemically characterized them, searching for a metabolically compatible solute responsible for generating the hydrostatic pressure. We grew spores of *M. grisea* in water drops on hydrophobic plastic membranes and allowed them to form appressoria. These generated full turgor over a period of 24–48 h. Gas-liquid chromatography showed that the most abundant solute in appressoria is glycerol.

Glycerol is generated rapidly during germination and germ-tube elongation (Fig. 1a). Here it probably contributes to plasma membrane biosynthesis during initial fungal growth⁷. Glycerol levels decrease at the onset of appressorium formation but rise sharply during turgor generation. This coincides with the collapse of the conidium and germ tube, and concentration of the cytoplasm within the unicellular appressorium (Fig. 1b). Intracellular glycerol concentration is considerably higher at this time (48 h), being contained in a small volume.

We estimate the mean internal volume

of an appressorium as $64 \mu\text{m}^3$ assuming that the cell is hemispherical with an internal radius of $3.1 \mu\text{m}$ ($n=100$, s.d.=0.5). The mean concentration of glycerol in an appressorium rose from 0.50 M at 24 h of development to 3.22 ± 0.40 M after 48 h. This is a conservative estimate because not all of the cell volume is available for glycerol accumulation. The osmotic potential generated by this concentration of glycerol would be -8.7 MPa at 20°C (assuming that glycerol is an ideal solute).

Glycerol solutions deviate from ideal at high concentrations. The maximum concentration of glycerol for which osmotic potential can be assayed psychro-

metrically⁸ is 2.0 M, but by extrapolation we estimate that 3.22 M glycerol would generate an osmotic potential of -5.8 MPa. This represents a minimum osmotic potential because living appressoria contain other solutes. The appressorium forms in a drop of water, so the turgor produced is at least 5.8 MPa.

To validate these estimates we incubated appressoria in a series of glycerol solutions of varying molarity and measured the frequency of cytorrhysis⁵ (cell collapse). The frequency was dependent on external glycerol concentration. A concentration of 1.75 M glycerol (-3.7 MPa) caused the collapse of 52% of appressoria (data not

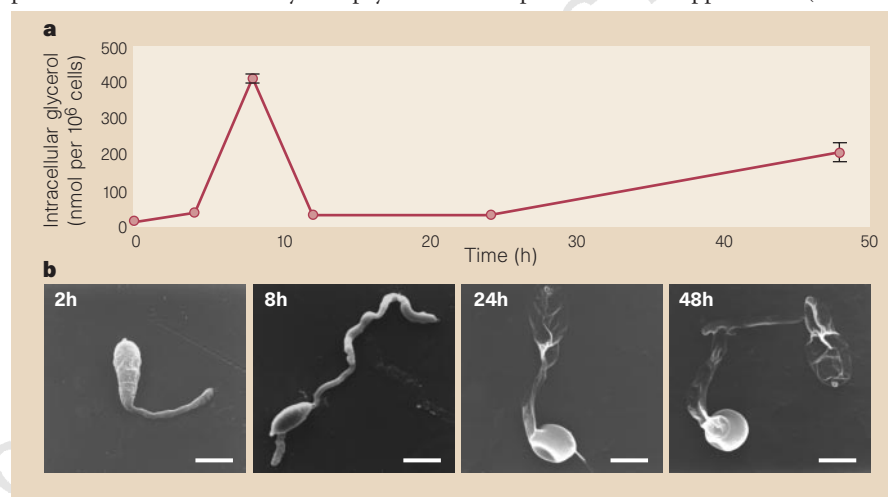


Figure 1 Intracellular glycerol increases during *M. grisea* appressorium turgor generation. **a**, Change in glycerol contents of extracts made from germinating conidia and developing appressoria with time, determined enzymatically using a glycerol-specific assay (Boehringer). **b**, Low-temperature scanning electron microscope images of developing appressoria at corresponding times. Appressoria were allowed to form in water drops on polytetrafluoroethylene (PTFE-Teflon, DuPont) membranes. As the appressorium develops turgor the conidium and germ tube collapse⁶. Cytoplasm is present only in the unicellular appressorium. Scale bars, $20 \mu\text{m}$.

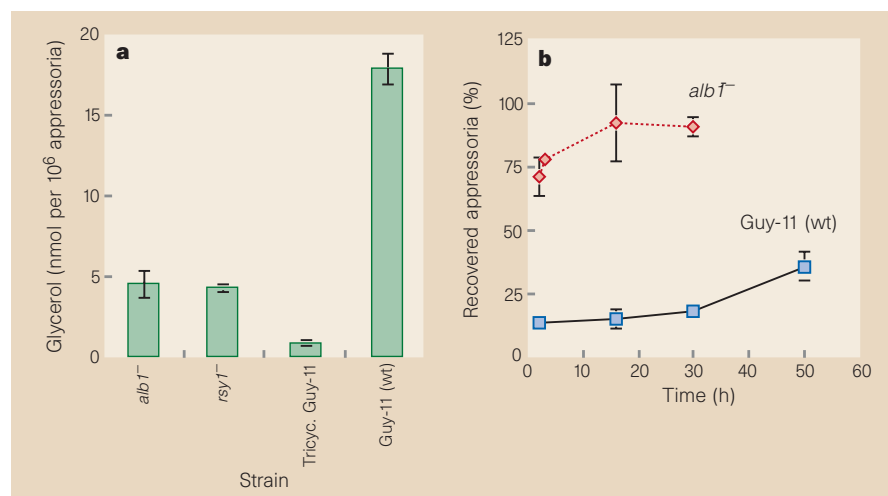


Figure 2 Non-melanized appressoria are permeable to glycerol. **a**, Intracellular glycerol levels in appressoria of *alb1⁻* and *rsy1⁻* mutants⁹, and of strain Guy-11 treated with tricyclazole, compared with wild-type untreated appressoria. We measured intracellular glycerol concentration after 24 h. **b**, We allowed appressoria to form in water for 24 h on PTFE membranes and then replaced the water with 4 M aqueous glycerol. We determined the rate of cytorrhysis and the proportion of appressoria that had recovered from cytorrhysis at each time point. Mean ($\pm$ s.d.) values from 100 appressoria in two independent experiments are represented. Melanized and non-melanized appressoria recovered fully after 60 min incubation in water (not shown).

shown). This supports the link between turgor and glycerol accumulation but also implies that the melanized wall is largely impermeable to glycerol.

Appressoria of *M. grisea* are heavily melanized^{9,10} and it has been shown genetically¹¹ that non-melanized appressoria fail to generate turgor and are non-pathogenic⁵. We found much lower levels of intracellular glycerol in appressoria from non-melanized strains carrying single gene mutations at *ALB1* and *RSY1* (Fig. 2), genes encoding enzymes required for dihydroxynaphthalene-melanin biosynthesis¹⁰. We found a similar reduction in glycerol accumulation after treatment of *M. grisea* with tricyclazole, a melanin biosynthesis inhibitor^{12,13} (Fig. 2a). Thus, melanin biosynthesis is required for efficient glycerol accumulation.

In cytorrhysis experiments we found that *alb1*⁻ mutant appressoria collapsed in hyperosmotic solutions of glycerol but quickly recovered (in under 1 min) and instead became plasmolysed⁵. This indicates that the non-melanized wall is permeable to glycerol and after initially causing cytorrhysis, glycerol diffuses through the cell wall and induces plasmolysis of the appressorial protoplast. This is in marked contrast to wild-type melanized appressoria which showed only limited recovery from cytorrhysis even after 48 h incubation in hyperosmotic glycerol (Fig. 2b). Maintenance of the enormous glycerol concentrations within appressoria is likely to be a consequence of the reduced permeability of melanized cell walls to glycerol preventing rapid leakage of the solute.

Several important plant pathogens form appressoria^{1,2} and although secretion of enzymes may aid cuticular degradation², mechanical infection of plant tissues is probably widespread. The infinite solubility and metabolic compatibility of glycerol therefore provides a simple and durable mechanism for plant infection which may be widely applied by pathogenic fungi.

Joke C. de Jong

Barbara J. McCormack

Nicholas Smirnoff

Nicholas J. Talbot

Department of Biological Sciences,

University of Exeter,

Washington Singer Laboratories,

Perry Road, Exeter EX4 4QG, UK

e-mail: N.J.Talbot@exeter.ac.uk

8. Koide, R. T., Robicheaux, R. H., Morse, S. R. & Smith, C. M. in *Plant Physiological Ecology: Field Methods and Instrumentation* (eds Pearcy, R. W. et al.) 161–183 (Chapman & Hall, London, 1989).

9. Howard, R. J. & Ferrari, M. A. *Exp. Mycol.* **13**, 403–418 (1989).

10. Chumley, F. G. & Valent, B. *Molec. Plant Microbe Interact.* **3**, 135–143 (1990).

11. Bourett, T. M. & Howard, R. J. *Can. J. Bot.* **68**, 329–342 (1990).

12. Woloshuk, C. P., Sisler, H. D. & Vigil, E. L. *Physiol. Plant Pathol.* **22**, 245–259 (1983).

13. Chida, T. & Sisler, H. D. *J. Pesticide Sci.* **12**, 49–55 (1987).

Reptile relationships turn turtle...

Turtles are so anatomically bizarre that their affinities with other reptiles remain contentious. Wilkinson *et al.*¹ showed how even the extensive morphological information amassed by Rieppel and deBraga^{2,3} provides only weak support for their view that turtles are advanced diapsid reptiles, rather than descendants of primitive anapsid reptiles, as conventionally thought. But there seem to be errors in Rieppel and deBraga's data matrix, many involving turtles or their putative 'anapsid' relatives^{4,5}. I have corrected and reanalysed the data (see 'Incorrectly coded characters', overleaf), and find that Rieppel and deBraga's data actually sup-

port, rather than challenge, the traditional view.

The modified data yield a tree where diapsid monophyly is restored and turtles are related to anapsid pareiasaurs (Fig. 1a). This result is consistent with other recent analyses^{4,5}. In particular, the tree is almost identical to that proposed in another detailed phylogenetic analysis of the entire Reptilia⁶ (Fig. 1b). The large impact of these apparently minor corrections to Rieppel and deBraga's data set is not surprising.

In their phylogeny (Fig. 1 in ref. 2) almost all of the characters interpreted as supporting turtle–diapsid affinities (those diagnosing clades 3 to 7) also occur in all or many anapsid 'parareptiles', or are absent (presumed reversed) in turtles. Thus, only a slight modification to the data caused turtles to shift from deep within diapsids (as lepidosaur relatives) to deep within parareptiles (as pareiasaur relatives).

My results, together with those of Wilkinson *et al.*¹, emphasize the importance of evaluating just how strongly a preferred tree is supported over alternatives. Rieppel and deBraga have identified a surprisingly strong phylogenetic signal linking turtles and advanced diapsids. However, the conventional views regarding

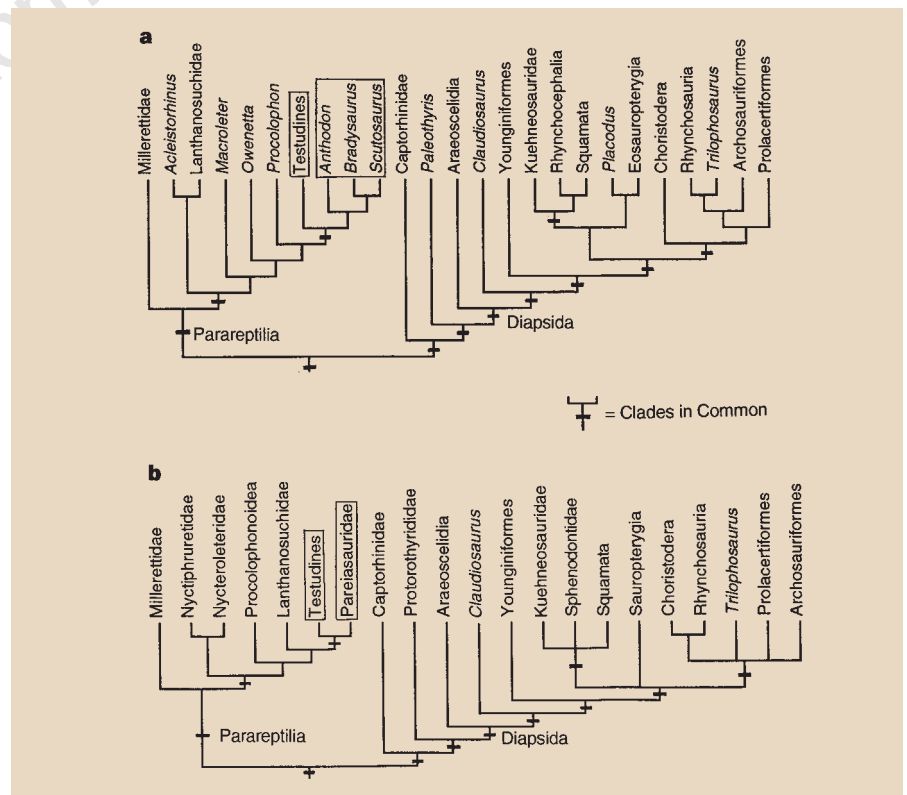


Figure 1 Alternative reptilian phylogenies. **a**, The single most parsimonious tree (length 763 steps, consistency index 0.511, retention index 0.695), based on an analysis of the data matrix of Rieppel and deBraga^{2,3}, after corrections detailed in footnote. Turtles (Testudines) are nested within 'anapsid' parareptiles, as the nearest relatives of pareiasaurs (*Anthodon*, *Scutosaurus*, *Bradysaurus*), and diapsid monophyly is restored. **b**, Phylogeny obtained from an independent study of higher-level reptile phylogeny⁶. Note the close similarity to **a**. Clades in common are indicated by horizontal bars. In both trees, diapsids are monophyletic and turtles are nested within parareptiles.

- Emmett, R. W. & Parberry, D. G. *Annu. Rev. Phytopathol.* **13**, 147–167 (1975).
- Mendgen, K., Hahn, M. & Deising, H. *Annu. Rev. Phytopathol.* **34**, 367–386 (1996).
- Talbot, N. J. *Trends Microbiol.* **3**, 9–16 (1995).
- Howard, R. J. & Valent, B. *Annu. Rev. Microbiol.* **50**, 491–512 (1996).
- Howard, R. J., Ferrari, M. A., Roach, D. H. & Money, N. P. *Proc. Natl Acad. Sci. USA* **88**, 11281–11284 (1991).
- Money, N. P. & Howard, R. J. *Fungal Genet. Biol.* **20**, 217–227 (1996).
- d'Enfert, C. & Fontaine, T. *Molec. Microbiol.* **24**, 203–216 (1997).

the anapsid derivation of turtles and the monophyly of diapsid reptiles need not yet be abandoned.

Michael S. Y. Lee

Zoology Building A08,
School of Biological Sciences,
University of Sydney, NSW 2006, Australia
e-mail: msylee@bio.usyd.edu.au

Incorrectly coded characters. I have provisionally accepted codings for other disputed characters (for example the acromion process²⁻³) where there is disagreement over the interpretation of homologies but room for debate. The data set was analysed using the same methods and outgroup rooting as refs 2 and 3. Characters are numbered as in ref. 3. **44–46**, Stapes unknown in *Bradysaurus* and *Anthodon*. **65**, Basal tubera on basioccipital and basisphenoid present in all pareiasaurs⁵. **70**, Rieppel and deBraga interpret the anterior braincase ossification in the primitive turtle *Proganochelys* as unequivocally a sphenethmoid, but then code turtles as lacking the element, instead of as polymorphic. **73**, Interpterygoid vacuity closed in *Bradysaurus* and slit-like in *Anthodon*⁷. **77, 82**, Transverse flange of pterygoid oriented anterolaterally, and jaw joint located anterior to occiput, in turtles (primitively)^{8,9}. **87**, Lateral shelf of surangular present in all pareiasaurs (for example, Figs 5–viii, 6 and Plate 33 of ref. 10). **103**, Cervical centra keeled in all pareiasaurs (for example, *Bradysaurus* BMNH R1971, *Scutosaurus* PIN 2005/1532–8, *Anthodon* BPI 1/548). **120, 121**, Two coracoids present in synsapsids¹¹. Rieppel and deBraga suggest that the acromion in turtles might represent a modified anterior coracoid. If so, the coracoid foramen in turtles lies totally within the coracoid homologues. Under their interpretation, turtles must be coded as equivocal for both coracoid number and coracoid foramen position. **124, 126, 127**, Humeral shaft short, supinator process absent, and ectepicondylar foramen present (primitively) in turtles⁸. **140**, Fourth trochanter absent in all pareiasaurs⁵. **150**, Well-defined articulation between astragalus and fourth distal tarsal present in *Scutosaurus* (for example PIN 2005/1877), condition unknown in *Bradysaurus* and *Anthodon*. **152**, First distal tarsal present (primitively) in turtles^{8,12}. **160**, Metapodials overlapping proximally in *Bradysaurus* (for example, plate 37 of ref. 13) and *Scutosaurus* (for example PIN 2005/1532), condition unknown in *Anthodon*. **164**, Long unguals primitively present in turtles^{8,9}.

1. Wilkinson, M., Thorley, J. & Benton, M. J. *Nature* **387**, 466 (1997).
2. Rieppel, O. & deBraga, M. *Nature* **384**, 453–455 (1996).
3. deBraga, M. & Rieppel, O. *Zool. J. Linn. Soc.* **120**, 281–354 (1997).
4. Laurin, M. & Reisz, R. R. *Zool. J. Linn. Soc.* **113**, 165–223 (1995).
5. Lee, M. S. Y. *Biol. Rev.* **70**, 459–547 (1995).
6. Lee, M. S. Y. *Zool. Scripta* (submitted).
7. Boonstra, L. D. *Ann. S. Afr. Mus.* **31**, 1–38 (1934).
8. Gaffney, E. S. *Bull. Am. Mus. Nat. Hist.* **194**, 1–263 (1990).
9. Rougier, G. W., de la Fuente, M. S. & Arcucci, A. B. *Science* **268**, 855–858 (1995).
10. Houghton, S. H. & Boonstra, L. D. *Ann. S. Afr. Mus.* **28**, 261–289 (1929).
11. Kemp, T. S. *Mammal-Like Reptiles and the Origin of Mammals* (Academic, London, 1982).
12. Zug, G. R. *Misc. Publ. Mus. Zool. Univ. Mich.* **142**, 1–98 (1971).
13. Houghton, S. H. & Boonstra, L. D. *Ann. S. Afr. Mus.* **28**, 297–367 (1930).

...and turn back again

The conventional view of evolutionary relationships within the living reptiles is that turtles are basal to the other groups. Their placement is based largely on the absence in turtles of temporal fenestrae, openings on either side of the skull involved in jaw muscle attachment, which all other groups of living reptiles and their descendants have. Rieppel and deBraga's recent analyses^{1,2} of fossil and living groups

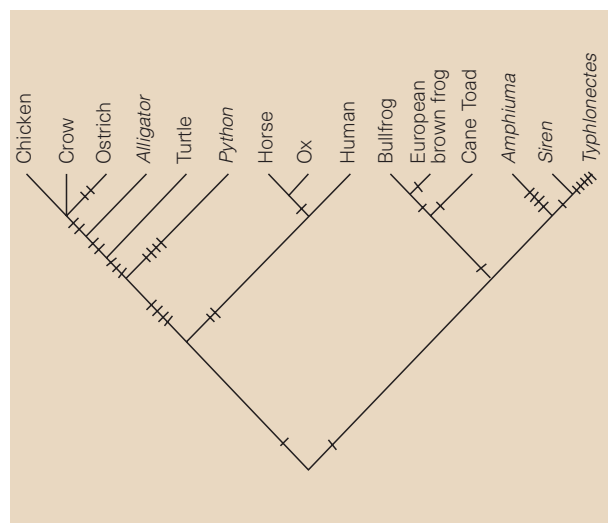


Figure 1 Strict consensus cladogram of the three shortest phylogenetic trees obtained by analysing tetrapod interrelationships on the basis of pancreatic polypeptide amino-acid sequences. Hash marks represent derived states for various amino acids. Those lying below a branch containing more than one taxon represent synapomorphies, each supporting the members of that branch. Similar marks on single terminal taxa are autapomorphies. The consistency index for each tree was 0.738.

of reptiles challenged this long-held conclusion, although their results have since been questioned³. We present a molecular analysis in support of Rieppel and deBraga's original conclusions.

Rieppel and deBraga¹ used 168 morphological characters, of which 13 (10 cranial and 3 postcranial traits) supported the placement of turtles within the ranks of reptiles that have twin temporal fenestrae on each side of the skull (the diapsid condition). On the basis of these 13 derived characters they concluded that there was robust support for considering turtles to be members of the crown group Diapsida. More recently, Wilkinson *et al.*³ reanalysed Rieppel and deBraga's data¹ and found that by using different topological constraints it required only three additional steps (773 as opposed to 770) to produce a tree which supports the original consensus of basal status for turtles. Wilkinson *et al.*³ acknowledged the difficulty of discerning the proper phylogenetic placement of turtles and suggested that molecular approaches would aid in determining the placement of turtles among the living reptiles.

Pancreatic polypeptide is a protein composed of 36 amino-acid residues and is a tetrapod product produced primarily in the islets of Langerhans. We determined the amino-acid sequence of pancreatic polypeptide for a representative of the turtles and compared it with published sequences for 14 additional tetrapod taxa. For phylogenetic analysis we used PAUP⁴. We examined representatives of five living taxa which have diapsid skulls or represent a derived condition from this ancestry (alligator, snake, and three species of birds). Nine additional taxa provided representatives of all other major groups of living tetrapods and include three orders of mammals and all three orders of living amphibians. The six amphibians represent tetrapods, a group that evolved before reptiles and were used in this study as the outgroup for rooting phylogenetic trees.

The 'branch and bound' method⁴ of analysis resulted in three equally parsimonious trees of 84 steps each. Of the 36 residues in pancreatic polypeptide, 24 were informative in establishing the branching pattern in Fig. 1. Of these residues, 15 were useful in defining the two main elements (mammals and reptiles) within the ingroup taxa that separate into two major clades shown to the left in Fig. 1.

All three trees produced the same two assemblages (mammals and reptiles) with the same nodal memberships within each group (consensus indices, 100%). The only node not resolved was that of the birds. The reptile clade was supported by four shared derived characters (Fig. 1). Based on the conventional view of turtles as anapsid reptiles, we had initially expected them to be basal to the other reptiles and birds, but this was not the case in any of the three trees. The branch order determining the placement of turtles was always immediately after that of snakes, which are clearly diapsid reptiles.

Our results were the same regardless of the topology constraints invoked during analysis. This branch of the tree is supported by three shared derived characters adding a molecular perspective congruent with the analysis by Rieppel and deBraga based on morphological grounds. Our results are therefore consistent with the recognition of turtles as diapsids and provide a focus for additional molecular considerations, which Rieppel and deBraga¹ and Wilkinson *et al.*³ encouraged.

James E. Platz*

J. Michael Conlon†

*Departments of Biology
and †Biomedical Sciences,
Creighton University, Omaha,
Nebraska 68178-0103, USA

1. Rieppel, O. & deBraga, M. *Nature* **384**, 453–455 (1996).
2. Rieppel, O. *Zoology* **98**, 298–308 (1995).
3. Wilkinson, M., Thorley, J. & Benton, J. B. *Nature* **387**, 466 (1997).
4. Swofford, D. L. *PAUP 3.1.1* (Smithsonian Inst., Washington DC, 1993).

The business of sustainability

Factor Four: Doubling Wealth, Halving Resource Use

by Ernst von Weizsäcker, Amory B. Lovins and L. Hunter Lovins

Earthscan: 1997. Pp. 332. £15.99

Charging Ahead: The Business of Renewable Energy and What It Means for America

by John J. Berger

Holt: 1997. Pp. 399. \$30

Robert Day

It would be easy for a business person or policy-maker to begin reading *Factor Four*, put it down halfway through and decide: "Yes, we're on the right track". What company is not constantly looking for greater efficiency? Which government is not encouraging its private sector to become more efficient? Achieving sustainable development solely through dramatic efficiency improvements appears at first glance to be the message of this book.

This impression, however, would be wrong. As the reader delves deeper into the text, perceptive and exciting ideas emerge for transforming our entire economic system towards sustainability. *Factor Four* provides a snapshot of the current state of the 'eco-efficiency revolution' but also demonstrates the limits we can achieve without making more radical changes to unleash the potential of the private sector.

At the beginning of the book, 50 examples of resource and energy efficiency drive home the point that there are many highly profitable opportunities for businesses, opportunities that actually build on each other. This is one of the more interesting ideas in the book — that instead of diminishing returns from eco-efficiency improvements, even greater opportunities are opened up for major process changes. A considerable improvement is usually found when the company realizes that it is providing a service and not a product. Utilities, for instance, are not selling electricity: they are keeping milk cold and houses warm. Is it necessary to pollute in order to do that?

Although the authors make it clear that they feel market failure is the rule and not the exception, they suggest policy changes and investment schemes that align market incentives with the needs of environmental protection, including 'polluter-pays' taxes, emissions trading and general regulatory reform. The authors conclude with a plea for greater investment by society in intangibles: "Markets were never meant to achieve community or integrity, beauty or justice, sustainability or sacredness", they write, "and they don't. If markets do something good for whales or wildness, for God or Gaia or

grandchildren, that is purely coincidental."

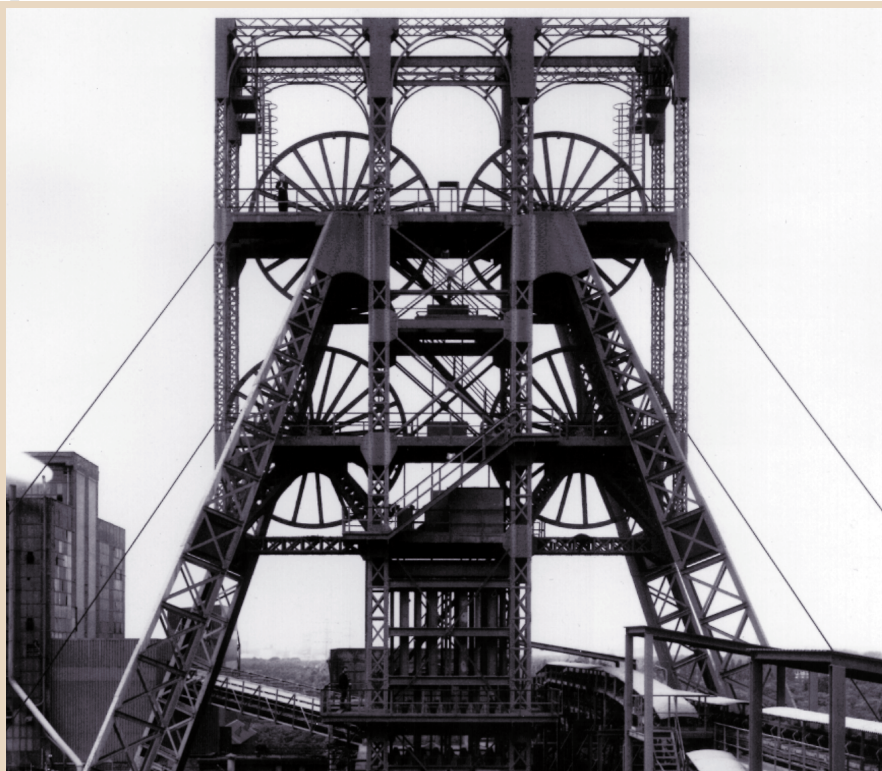
Such ideas are important because the authors explicitly acknowledge the need for business to make a profit. Parts of the book, however, are a little confusing. Much of the policy discussion, for instance, focuses on the experiences of utilities with energy savings (or 'negawatts') markets, regulators and deregulation. The lessons drawn are interesting and useful, but readers unfamiliar with the peculiarities of utility regulation in the United States may find more use in such exciting ideas such as the 'golden carrot' awards — monetary awards given to the first company that not only invents a super-efficient household appliance but also successfully brings it to the marketplace. Through such stories, the reader can quickly see how small, inexpensive incentives can 'jujitsu flip' market forces to drive eco-efficiency.

Although the book asserts that business can use the eco-efficiency idea to achieve deeper social and environmental goals, its roadmap for doing so is somewhat vague. We cannot back our way into a sustainable future by simply doing what we do more efficiently. For example, the authors point out that while CAFE (fuel efficiency) standards for vehicles have lowered fuel use per mile in

the United States, Americans now drive more miles per person. We cannot achieve sustainable development without real change in our economic patterns, not just process improvements.

One way to think about the difference between eco-efficiency and sustainable development is in the context of public goods. Public goods such as infrastructure, clean air and water, social culture and security are used productively by all businesses, big or small. Eco-efficiency is all about minimizing our negative impact on the sources of these public goods: the environment, society and people. Sustainable development, on the other hand, requires us to contribute to these public goods. To achieve sustainable development, we must create systems that allow businesses to profit from the production of public goods.

This is not as impossible as it seems. Occasionally, the collective action of the market itself may be enough to provide sufficient incentives. There are opportunities right now for businesses that look hard enough. For instance, telecommunications companies may find it advantageous to build a public wireless communications network — because it will mean that more people use their product — even though they are pro-



On the shoulders of giants

Bernd and Hilla Becher see themselves as 'industrial archaeologists', capturing on film the "delicate giants" that stand over the shaft

entrances to mines in Europe and North America. This picture is one of 180 in their album *Mineheads* (MIT Press, \$75, £49.95).

viding infrastructure that could also be used by their competitors' products. Government, as the custodian of public goods, can also help. Its laws and regulations must be designed to reward businesses for providing public goods. In effect, government is 'outsourcing' its job through incentive-based regulation. To their credit, the authors of *Factor Four* take on these and other difficult issues, and they recognize the need for greater change beyond 'eco-efficiency'.

Charging Ahead gives a good example of this principle. This book is a definitive description of the US renewable energy industry and an argument for regulatory change to support it. The author has apparently talked to almost everyone involved in the industry, and he includes many fascinating case studies. Although the book is focused on industry in the United States, its value extends beyond that country. The book gives an overview of the industry and the technologies involved. Although it is a snapshot of what was state-of-the-art in 1996, it is comprehensive and clearly written. Even readers with little or no previous knowledge of the subject will be able to follow the author. The case studies themselves — with their personal stories of struggle and triumph — yield valuable and basic lessons on entrepreneurship that could be useful in a business-school classroom. Finally, the book provides resources for researchers, including a directory of organizations active in the field.

Businesses today that strive for eco-efficiency often face a trade-off between materials and energy efficiency. For instance, they can use processes that require fewer raw material inputs but are more energy-intensive. This results in more greenhouse-gas emissions and other pollution. Many of the examples given in *Factor Four* demonstrate this trade-off, a large obstacle to achieving sustainable development through improvements in efficiency.

The root of the problem is that we obtain most of our energy from physical materials such as fossil fuels. The result is that these efficiency improvements simply trade one form of material for another. What if there were a source of energy that did not require pollution or massive materials use? Then we could have energy-intensive industrial processes with no environmental trade-off for high material efficiency. Then both business and society could meet their needs.

Charging Ahead declares that one such source of energy is the Sun. Unfortunately, official and unofficial government subsidies for fossil fuels and nuclear power make it appear that photovoltaics, biomass, wind turbines and other renewable energies derived from the Sun are not as economically attractive. But the fact is that the true cost to society of a gallon of gas is much more than what we pay at the pump. Nevertheless, the

author argues that technology is reaching a level where these energy alternatives will soon be inexpensive enough to begin taking over the market. He suggests that the United States should not only eliminate the subsidies given to fossil and nuclear fuels but that it should also begin investing in domestic research in renewables. Otherwise the United States will fall behind in the coming energy revolution. Specific policy recommendations are given.

Charging Ahead provides good examples of profiting through sustainable development. The author argues that the United States can contribute to human development and also gain economically by encouraging business to sell renewable energy technology to the developing world. He describes several businesses that are already doing this on a small scale.

When read together with *Factor Four* the message of a radical, profitable and sustainable change comes through loud and clear. Both books provide a challenge to industry and policy-maker to look for the business opportunities in sustainable development. □

Robert Day is at the Sustainable Enterprise Initiative, World Resources Institute, 1709 New York Avenue NW, Washington DC 20006, USA.

Emergent learning

Connected Knowledge: Science, Philosophy, and Education

by Alan Cromer

Oxford University Press: 1997. Pp. 221. \$25, £19.99

Eugenie C. Scott

Alan Cromer is a man with a mission. A self-described "optimistic know-it-all", he wants US science education to shape up and abandon constructivism and other trends that he feels are not only failing to educate young Americans but are also mis-educating them. As in his earlier book, *Uncommon Sense* (Oxford University Press, 1993), he presents science as a non-intuitive way of knowing about phenomena whose causes are not obvious. To learn how the world works requires systematic introduction of principles that build upon one another. Scientific understanding is based on feedback between theory and experience, spiralling up to a more complete understanding of nature. He blends this philosophy of science with the teaching of science, taking us through quantum physics, the nature of the social sciences, his personal theory of human social organization, a history of education, some idiosyncratic views of learning theory, an even more idiosyncratic commentary on genetics, race, class and IQ, and finally his recommendations for reorganizing US science education.

There is no doubting Cromer's passion

and dedication. He has spent several years with Project SEED, an innovative teacher-training programme, as well as developing programmes for teaching basic scientific literacy to prison inmates. He has thought carefully about what high-school graduates should know and about how to design a curriculum to achieve these objectives. His advice flies in the face of some cherished values in US education — which alone recommends the book.

Constructivism, promoted in all reform guidelines on US science education, comes in for a sound thrashing. In constructivist learning, students 'construct' knowledge of science by 'discovering' principles, especially 'learning by exploring'. This is thought to result in deeper and more complete understanding. Critical thinking is taught through 'open-ended problems' with either no clear answer or many possible answers.

Wrong reasoning, says Cromer. To learn even the simplest scientific principles (he takes most of his examples from physics) requires the direction of a teacher. "In mucking about randomly, a student learns as little as a mouse does while meandering about the maze on its first trial. Only when the student reaches a goal, such as getting an experiment to agree with an equation, does the whole enterprise begin to make any sense.... An experienced science teacher knows that some detours are so wasteful of time and energy that students should be warned against them, whereas other byways might be left for the students to explore... it may seem far-fetched to compare a student doing a physics experiment with a mouse running through a maze, but only to someone who has never taken a physics laboratory course."

But the term 'constructivism' hides much variation, and I have seen some classroom teachers leading students to understanding in precisely the way Cromer recommends — and calling it constructivism. Some constructivist approaches cheerfully if mindlessly exhort the teacher to "accept all answers" instead of reminding them that the point of the exercise is to help the student to understand some principle or other. But some uses of constructivism, such as in pre-assessing a student's understanding (or misunderstanding) of a topic, are certainly worthwhile. There is a place for letting students explore, as Cromer would agree. But for a student to understand either basic principles of science or how science works, the guidance of a teacher is necessary because "without knowledgeable guidance from their teacher, students are truly like mice in a maze. Each will arrive at his own version of the goal with his own set of errors and misconceptions." What is dismal about US education is that most teachers do not understand enough about basic science to be able to supervise such explorations properly, whether called constructivist or something else.



Old companions who learn new tricks

Hunting dogs take to the water with their Micmac owners in nineteenth-century Canada. In *A History of Dogs in the Early Americas* (Yale University Press, \$27.50, £21), Marion Schwartz

tells how these versatile companions have coexisted with humans for 12,000 years, whether reared for food, used in hunting, sent to war or revered as guides to the afterlife.

So what is Cromer's solution? We must, he says, agree on what high-school graduates should know about science and then develop a coherent curriculum to produce students who have this knowledge. Cromer criticizes the US National Science Education Standards for not accomplishing this task and being laden with educational gobbledegook. A *de facto* set of ideas and skills already exists, he says, in the 'General Educational Development' test, or GED, a rigorous seven-hour high-school equivalency test of language skills, social studies science and mathematics given to adults. A pass in GED or a high-school diploma is required in the United States to attend college or technical school or to apply for most jobs. A novel idea is to have all ninth-graders take GED before they can go on to the final three years of high school — or leave school or go into a training programme. "It is vitally important that there be a meaningful intermediate certificate to provide young people with an honorable way to leave school after ninth or tenth grade. The drive to push everyone through twelve years of academic study has made 'drop outs' of those who are unable or unwilling to do so."

To ensure all students get at least to the GED level will require another unfashionable idea — ability grouping, in which students are grouped by their ability to perform certain tasks. Cromer does not intend this to be 'tracking', where students are permanently assigned to high, medium or low IQ groups, but a looser, less-permanent grouping that students can move in and out of as their skills and knowledge improve. He

believes ability grouping is especially important if low-achieving students are to meet the minimal GED-type standards, because these students need special attention to develop even the most basic understanding.

Cromer enjoys the role of curmudgeon, and the forceful way in which he writes cannot help but engage the reader. (Speaking of criticisms of intelligence testing, he growls: "There are few educators who know enough arithmetic to balance a checkbook, let alone understand a multivariate logistic regression analysis".) But this leads to the occasional overstatement that frustrates or annoys. His physics is better than his social science and history. A chapter explaining why the uncertainty and indeterminacy of quantum mechanics makes the visible world in which we live certain and determined is an excellent antidote to postmodernists' claims about the lack of objective reality and the supposed inability of science to explain it. His discussion of intelligence is better than his discussion of race: as a physical anthropologist, I was dismayed by his confusion of the concept of equality with that of identicalness (the former social and legal construct is independent of the latter biological one). He combines the principles of natural selection with observations of animal and human behaviour to produce a new theory of human social organization. Here the yin of hierarchy and loyalty was selected with the yang of individualism and rebelliousness as adaptive traits in early human social environments. Although I am generally sympathetic to sociobiological and evolutionary

approaches, I did not find the discussion fully persuasive.

But these are minor glitches in what is certainly a stimulating and thought-provoking book. Although the hats on the good guys and the bad guys are perhaps both whiter and blacker than in reality, one should definitely consider Cromer's analysis. There is a lot to be said for systematically teaching science from the part to the whole. □

Eugenie C. Scott is at the National Center for Science Education, 925 Kearney Street, El Cerrito, California 94530-2810, USA. E-mail: scott@natcensied.org

Fins, legs, fins

Ancient Marine Reptiles

edited by Jack M. Callaway and Elizabeth L. Nicholls

Academic: 1997. Pp. 501. \$64.95, £49.95

Michael W. Caldwell

This book hails itself as the long overdue revision of Samuel Williston's *Water Reptiles of the Past and Present* (1914). The claim invites comparison. Williston wrote his book by himself; 28 authors, contributing 17 chapters and 6 introductions, have produced *Ancient Marine Reptiles*. Williston's book was devoted to a taxonomic and biological review of marine reptiles, and not one page described or named a new taxon; four chapters in *Ancient Marine Reptiles* are purely descriptive, serving only to give names to, or to revise, a single genus or species.

If there is a critical flaw in this edited volume, if there is one feature where it drifts from the tradition of Williston's book, this is it. Simply put, these four chapters should have been published as journal contributions, leaving more space for the remaining 13 synthetic chapters. To my mind, an up-to-date assessment of a particular field should focus on the contrast between observation and theory; innovation is found in the distillation and synthesis of disparate data points and the resulting generation of new questions. It is a pity that there was not more room in this volume to address these contrasts in depth.

Nevertheless, I applaud this volume. Most chapters are well written and pertinent. The figures are informative and the references accurate. Highlights include the chapters by McGowan and by Motani, which provide excellent reviews and interpretations of new faunas and data sets; Rieppel's chapter on Triassic sauropterygians, even though it barely reviews his recent mountain of revisionary publications; and Bell's chapter on mosasaur phylogeny, which unfortunately only scratches the surface of his PhD thesis, with no discussion of the implications of his character analysis.

The last four chapters, with a lengthy introduction by Massare, reveal the main

palaeobiological issues underlying the study of ancient marine reptiles. Did fins evolve to legs and back to fins again? And what was more important, homoplasy or homology? Selection or constraint? Unique morphologies or modifications of the old? Marine reptiles are a model system for posing a broad range of questions about the evolution of tetrapods in aquatic environments. We often consider cetaceans or pinnipeds in this light, yet forget that ichthyosaurs were 'dolphin-like' about 190 million years before the first proto-dolphin even thought of going for a swim. □

Michael W. Caldwell is in the Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9.

Axioms for biology

Foundations of Biophilosophy

by Martin Mahner and Mario Bunge
Springer: 1997. Pp. 423. \$54, £36.50

Paul E. Griffiths

The logical empiricist consensus that existed in the philosophy of science until the 1960s held that the ideal statement of a scientific theory would be a formal axiom system of the kind found in mathematics. The biologist J. H. Woodger is remembered for his attempts to make biology live up to this ideal. For reasons too complex and varied to explore here, few philosophers of science now try to make scientific theories axiomatic.

It is therefore surprising to find the biologist Martin Mahner and the philosopher Mario Bunge formulating their general account of biology using the logical apparatus of predicate calculus and set theory, complete with axioms, definitions and corollaries. Their aim appears twofold: to ensure consistency between their views on diverse topics and to ensure that their biological views are as tightly constrained as possible by the metaphysics that occupies the first third of their book. This introductory section takes a stand on most of the basic questions of metaphysics and epistemology. Topics covered include basic ontology (the nature of events, properties and things), the interpretation of the probability calculus and the nature of truth and evidence. Philosophers will be frustrated at the speed with which Mahner and Bunge dismiss alternative positions, and biologists should beware of taking their often strongly stated views to represent a philosophical consensus. However, it is hard to see how these problems could be avoided without turning this into a book on metaphysics and epistemology, rather than biology. Bunge has argued at length for his views in many other works.

Perhaps the biggest surprise is that there is no discussion of the reduction of Mendelian to molecular genetics. This is the topic in the philosophy of biology most

influenced by the axiomatic method. The original idea of 'theory reduction' was that the axioms of one, formalized theory should be derived logically from those of another, thus showing that nothing is lost when the former is replaced by the latter. This omission is more surprising because the coverage of the book is otherwise excellent.

There are two further surprises in Mahner and Bunge's biophilosophy. First, they reject 'population thinking': evolution is not about ensembles of individuals but about types of organism. They argue that these types, which include species, are in some sense 'natural kinds'. They strongly reject the consensus view, deriving from David L. Hull and Michael Ghiselin, that taxa are ontologically akin to individual objects such as nations rather than to natural kinds such as gold. The second surprise is the authors' radical approach to developmental biology: development is not guided by a genetic program. Instead, Mahner and Bunge call for a synthesis between the 'structuralist' approach to development, which seeks emergent laws of complex biological systems, and the 'constructionist' view that the control of development cannot be localized in one material cause.

Having adopted this radical perspective, Mahner and Bunge attack existing structuralists and constructionists in a manner worthy of Trotskyite splinter politics. The structuralism of Brian Goodwin must be rescued from "holism, (crypto)idealism and subjectivism" by the adoption of Mahner and Bunge's metaphysical scheme. My own advocacy of constructionism with the biologist Russell Gray "must be judged as an utter failure" because of a "severely flawed ontology" which I share with A. N. Whitehead and Hull.

Mahner and Bunge convict most existing biological theorists of basic metaphysical errors, something they hope to avoid through their formal, axiomatic method. For example, Theodosius Dobzhansky characterized evolution as change in the genetic composition of populations owing to "altered interactions with their environment". The authors may be right that it is strictly organisms rather than populations that interact with the environment, but I remain unconvinced that Dobzhansky's slip shows "how easily habits of speech may obfuscate clarity and proper theorising".

This tendency to over-diagnose fundamental metaphysical confusions and to deduce absurd apparent consequences reduces the usefulness of this book as a text for students, despite its admirably wide coverage of the subject. The axiomatic framework and use of logical symbolism will also be unattractive to students, especially those in biology. □

Paul E. Griffiths is in the Department of Philosophy, University of Otago, PO Box 56, Dunedin, New Zealand.

When the rains come

Climates of South Asia

by G. B. Pant and K. Rupa Kumar
Wiley: 1997. Pp. 320. £75, \$130

Julia Slingo

The climate of South Asia is dominated by the monsoon, which returns with remarkable regularity each summer and provides the rainfall needed to sustain more than 60 per cent of the world's population. The vastness of the Asian continent and the unique configuration of the East African Highlands and the Tibetan Plateau mean that the Indian summer monsoon is the most vigorous and influential of all the monsoon circulations. Increasingly, the monsoon is being seen as an important player in the global climate.

It is fortuitous that the first book detailing the climatology of South Asia should be published on the fiftieth anniversary of India's independence. Without the influence of the British Raj and scientists such as Blanford, who in 1886 recognized "the paramount importance of knowledge of distribution of rainfall in space and time", the comprehensive database that provides the core of this book might not have been realized.

The book brings together the vast literature and data resources of India and its neighbours (Pakistan, Bangladesh, Nepal and Sri Lanka), providing a comprehensive description of a wide range of topics, from extreme weather events to the environmental impact of climate change. It is generally well written, although unfortunately the authors had not updated much of the material to include the most recent observations. For those seeking an understanding of the physical mechanisms involved in the climatology of South Asia, its weather systems and its year-to-year variability, this book provides little in the way of answers. That said, it is a valuable teaching aid, is likely to appeal to the more casual reader and is an excellent reference book for dipping into. □

Julia Slingo is at the Centre for Global Atmospheric Modelling, University of Reading, Earley Gate, Reading RG6 6BB, UK.

Related books

Antarctic Meteorology and Climatology by J. C. King and J. Turner. CUP, £55, \$90.

El Niño Southern Oscillation and Climatic Variability edited by R. Allan, J. Lindesay and D. Parker. CSIRO, \$110.

The Weather of Britain by Rob Stirling. Revised pbk. Giles de la Mare, £19.99.

Watching the Weather by John and Mary Gribbin. Constable, £7.95 (pbk).

The Weather Book: An Easy-to-Understand Guide to the USA's Weather by Jack Williams. Revised pbk. Vintage, \$20.

Crystal structure of the nucleosome core particle at 2.8 Å resolution

Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent & Timothy J. Richmond

Institut für Molekularbiologie und Biophysik ETHZ, ETH-Hönggerberg, CH-8093 Zürich, Switzerland

The X-ray crystal structure of the nucleosome core particle of chromatin shows in atomic detail how the histone protein octamer is assembled and how 146 base pairs of DNA are organized into a superhelix around it. Both histone/histone and histone/DNA interactions depend on the histone fold domains and additional, well ordered structure elements extending from this motif. Histone amino-terminal tails pass over and between the gyres of the DNA superhelix to contact neighbouring particles. The lack of uniformity between multiple histone/DNA-binding sites causes the DNA to deviate from ideal superhelix geometry.

DNA in chromatin is organized in arrays of nucleosomes¹. Two copies of each histone protein, H2A, H2B, H3 and H4, are assembled into an octamer that has 145–147 base pairs (bp) of DNA wrapped around it to form a nucleosome core (of relative molecular mass 206K). This highly conserved nucleoprotein complex occurs essentially every 200 ± 40 bp throughout all eukaryotic genomes². The repeating nucleosome cores further assemble into higher-order structures which are stabilized by the linker histone H1 and these compact linear DNA overall by a factor of 30–40. The nucleosome (nucleosome core, linker DNA and H1) thereby shapes the DNA molecule both at the atomic level through DNA bending, and on the much larger scale of genes by forming higher-order helices³. The nucleosome, in its role as the principal packaging element of DNA within the nucleus, is the primary determinant of DNA accessibility.

The physical properties of nucleosomes depend on solution conditions such as ionic strength, divalent-ion concentration, and on histone-modification state⁴. For example, the variability of nucleosome spacing within *in vitro*-assembled arrays demonstrates the underlying electrostatic component of the interaction of DNA with the basic histone proteins⁵. The characteristics of an individual nucleosome, however, depend on the actual DNA sequence incorporated, and this can have functional significance for specific gene promoters⁶. Positioning of nucleosomes along the DNA molecule is in part governed by DNA sequence preferences, but once formed, a nucleosome may not be entirely fixed in position⁷. The statistical preference observed for the DNA minor groove to face the histone octamer at (A + T)-rich sequences indicates that certain sequences will impart a rotational orientation to the DNA double helix when bound in a nucleosome⁸. The variation in this preference along the length of nucleosomal DNA and the intrinsic sequence-dependent bendability of DNA implies that strong translational settings can also occur. Both types of positioning have been detected with base-pair accuracy in the mouse mammary-tumour virus long terminal repeat⁹.

Recombination, replication, mitotic condensation and transcription involve the chromatin substrate and are thus affected by its structure. The generally repressive nature of chromatin structure has long been appreciated in transcription regulation¹⁰. Chromatin organization can facilitate the activation of specific genes^{11,12}. Evidence supporting two themes is accumulating. First, histones collaborate with transcription factors to provide for their own removal or structural modification, resulting in gene derepression. This may occur by stepwise invasion of DNA-binding proteins entering at the terminal segments of the nucleosomal DNA¹³. This mechanism is apparently insufficient in general, however, because several ATP-dependent chromatin remodelling factors have been

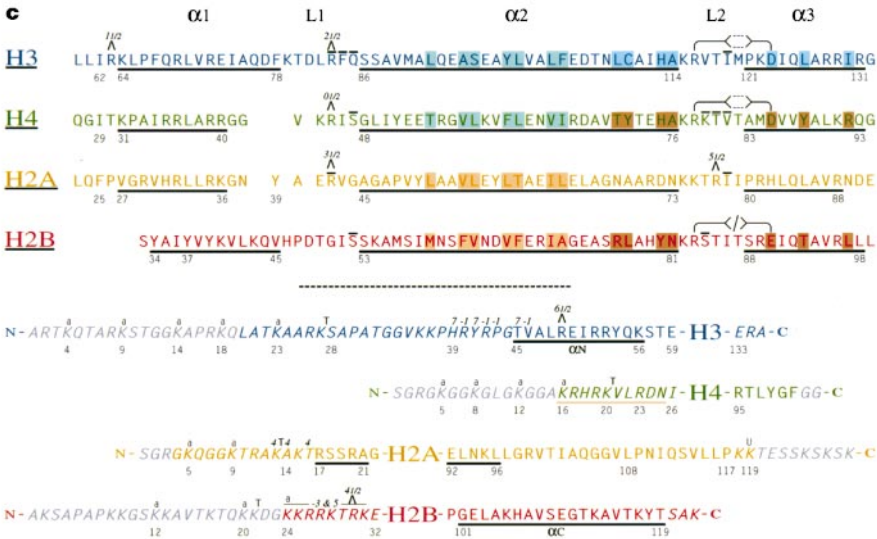
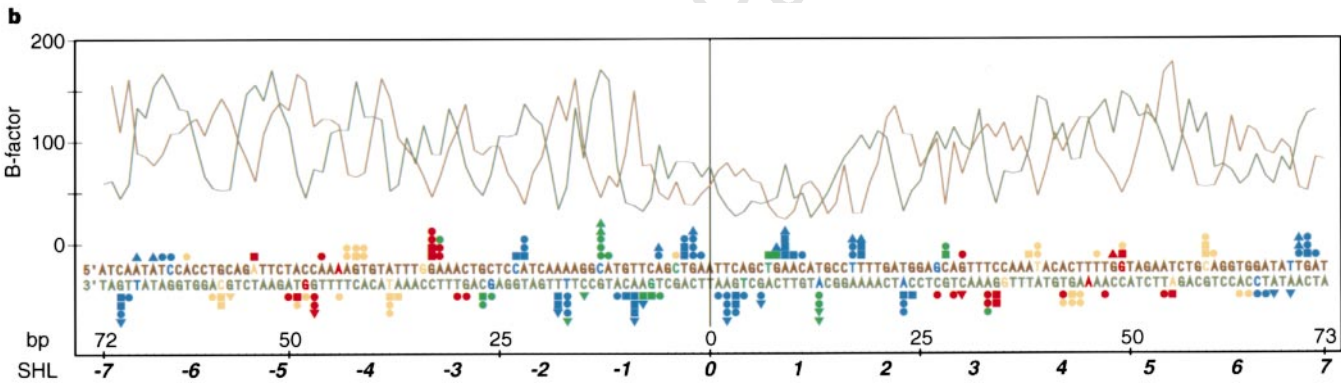
discovered that cause the core histones partially to lose or alter their grip on DNA¹⁴. Second, the bending and supercoiling of DNA on a nucleosome can promote binding of transcription factors and augment interactions between different factors^{15,16}.

The structure of a nucleosome core particle, crystallized under physiologically relevant ionic conditions¹⁷, was previously solved by X-ray crystallography at 7 Å resolution¹⁸. This low-resolution structure confirmed the disc-like shape of the particle deduced from earlier studies¹⁹ which results from the flat, left-handed superhelical arrangement of the DNA. The path of the DNA superhelix was irregular, bending tightly at several positions and being nearly straight in others. Although detailed side-chain information or atomic coordinates are not available, the re-interpreted, high-salt, histone octamer crystal structure, for which the histone fold was first described²⁰, allowed more accurate assignment of the secondary structure of the histone proteins in the core particle at low resolution²¹.

We have determined the high-resolution X-ray structure of the nucleosome core particle from new crystals which have all components individually made in bacteria and assembled after purification. These crystals, containing DNA of defined sequence and histones lacking post-translational modifications^{22,23}, diffract anisotropically to between 1.8–2.2 Å. Recombinant histone proteins enabled heavy-atom derivatives to be prepared through the substitution of cysteine at many sites. Use of an ESRF undulator beam line allowed accurate measurement of the weak diffraction intensities from these crystals and yielded a high-quality electron density map. The structure presented here is currently refined to 2.8 Å resolution, and includes an entire 146-bp DNA molecule and over 80% of the eight histone chains. Some portions of the N- and C-terminal histone-tail regions extending outside the confines of a single particle have not been included in the atomic model.

Organization of the nucleosome core particle

The 146 bp of DNA are wrapped around the histone octamer in 1.65 turns of a flat, left-handed superhelix (Fig. 1a). Although the human α -satellite DNA used was made palindromic to accommodate perfect twofold symmetry, the sequence in fact binds the histone octamer with a central base pair at the particle pseudo-twofold axis so that the DNA is divided into 73- and 72-bp halves, with one base pair falling on the dyad. We define the rotational orientation of the DNA double helix relative to the central base pair (superhelix location zero, or SHL0, where the major groove faces the octamer), and for each successive turn, the location number increases in the 73-bp half up to SHL + 7, and decreases in the 72-bp half down to SHL – 7 (Fig. 1d; only the 73-bp half of the DNA superhelix and associated histone proteins are shown for clarity). The superhelix



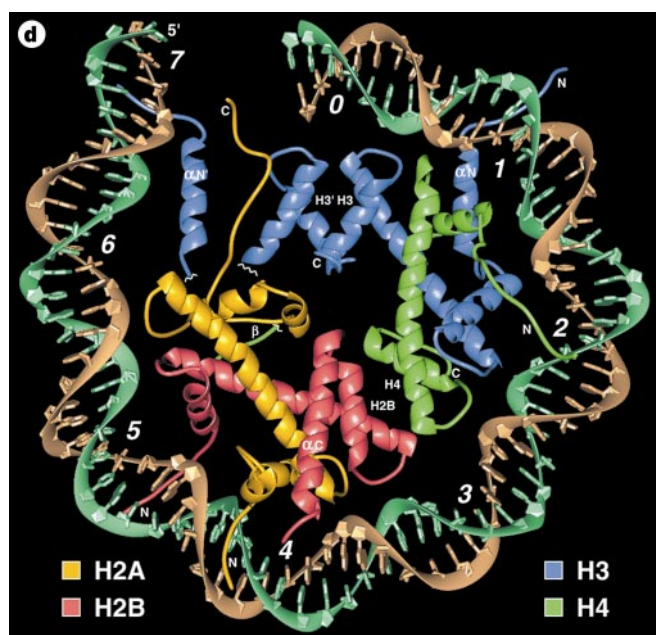


Figure 1 a, Nucleosome core particle: ribbon traces for the 146-bp DNA phosphodiester backbones (brown and turquoise) and eight histone protein main chains (blue: H3; green: H4; yellow: H2A; red: H2B). The views are down the DNA superhelix axis for the left particle and perpendicular to it for the right particle. For both particles, the pseudo-twofold axis is aligned vertically with the DNA centre at the top. **b**, DNA phosphate *B*-factors versus base pair. The sequence of the DNA used is shown with corresponding *B*-factors ($\text{\AA}^2$) plotted for the 5' phosphate group of each base. The contacts of the DNA phosphodiester chains with the histones are indicated: squares for main-chain hydrogen bonds; circles for side-chain hydrogen bonds, and triangles for hydrophobic bonds. The bases coloured blue, green, red, and yellow indicate close proximity to an arginine side chain inserted into the minor groove. The DNA phosphate groups have high mobility or are disordered when not contacted by the histones. **c**, Core histone sequences. The histone fold regions for H3, H4, H2A and H2B are aligned on the basis of their structures and labelled on the left (top). The α -helix ($\alpha 1$, $\alpha 2$, $\alpha 3$) and loop (L1, L2) secondary structural elements are labelled; α -helices are underlined. Δ designates an arginine side chain that is inserted into the DNA minor groove at the indicated SHL. The overlying lines indicate β -strand hydrogen bonds between L1 and L2 loops. The side chains indicate the positions of buried arginine-aspartate pairs for H3-H4 and the lack of a homologous pair for H2A-H2B. The histone-fold extensions and tail (italics) regions are shown with the core histone name representing the histone fold (bottom). Grey-shaded sequences were not included in the atomic model. Sites of *in vivo* acetylation are indicated with a letter 'a'. Trypsin cleavage sites that produce a 'tailless' core particle are denoted by T (ref. 50). The site of ubiquitination is marked U (ref. 51). SHL labels above the sequence indicate contacts with DNA at those points. The H4 sequence at amino acids 16-25 and underlined in orange makes a clear interparticle contact with a H2A-H2B dimer. Some interaction sites are highlighted: (1) H3-H4 $\alpha 2$ - $\alpha 2$ (blue-green); (2) H2A-H2B $\alpha 2$ - $\alpha 2$ (orange); (3) H3 4-helix bundle (blue); (4) H4-H2B 4-helix bundle (brown). **d**, Nucleosome core particle: 73-bp half. The view is down the superhelix axis with the pseudodyad axis aligned vertically. The central base pair through which the dyad passes is above the SHL0 label, 0 (SHL, superhelix axis location). Each SHL label represents one further DNA double helix turn from SHL0 (1-7). The complete histone proteins primarily associated with the 73-bp superhelix half are shown (interparticle tail regions are not shown). The two copies of each histone pair are distinguished as unprimed and primed copies, where the histone fold of the unprimed copy is primarily associated with the 73-bp DNA half and the primed copy with the 72-bp half. The 4-helix bundles are labelled as H3' H3 and H2B H4; histone-fold extensions of H3 and H2B are labelled as $\alpha N'$, αN and αC , respectively; the interface between the H2A docking domain and the H4 C terminus as β ; and N- and C-terminal tail regions as N or C.

has an average diameter of $41.8 \text{ \AA}$ but is not uniformly bent, having maximum curvature and radius at $\text{SHL} \pm 1.5$ and $\text{SHL} \pm 4-5$. The difference in the relative mobility of the segments of the DNA phosphate chains facing the solvent compared with those bound to the histones is extreme compared with other protein/DNA complexes, as revealed by the 2-3-fold variation in crystallographic *B*-factors (Fig. 1b), and may explain the relatively weak diffraction intensity from the crystals.

The protein octamer is divided into four 'histone-fold' dimers defined by H3-H4 and H2A-H2B histone pairs (Fig. 1c). The two H3-H4 pairs interact through a 4-helix bundle formed only from H3 and H3' histone folds to define the H3-H4 tetramer. Each H2A-H2B pair interacts with the tetramer through a second, homologous 4-helix bundle between H2B and H4 histone folds. The histone-fold regions of the H3-H4 tetramer bind to the centre of the DNA covering $\text{SHL} - 3$ to $+3$, whereas those of the H2A-H2B dimers bind from -6 to -3 and $+3$ to $+6$ (Fig. 1b, d). Further α -helices and coil elements (such as H3- αN and H2B- αC) extend from the histone-fold regions and are also an integral part of the core protein within the confines of the DNA superhelix. Beyond these histone-fold extensions, the histone N- and C-terminal tail sequences, which make up about 28% of the mass of the core histone proteins, are seen over about one-third of their total length in the electron density map. The tails of H3 and H2B pass through channels in the DNA superhelix created by two juxtaposed minor grooves. One H4 tail segment makes a strong interparticle connection, perhaps relevant to the higher-order structure of nucleosomes.

Histone-fold pairs

The central histone-fold domains of all four core histone proteins share a highly similar structural motif constructed from three α -helices connected by two loops, L1 and L2, denoted as $\alpha 1$ -L1- $\alpha 2$ -L2- $\alpha 3$ (Fig. 1c). These regions form crescent-shaped heterodimers in the pairings H3-H4 (Fig. 2a) and H2A-H2B (Fig. 2b) and bind 2.5 turns of DNA double helix, which arcs around them along their long axes to generate a 140° bend. The histone-fold motif is related to its counterpart in a pair by pseudo-twofold symmetry, with the symmetry axis passing between the two long, 8-turn $\alpha 2$ helices at their crossover. Their antiparallel arrangement places the L1 loop of one in juxtaposition with the L2 loop of the other. As the contact surfaces are offset towards the N terminus by one helical turn, the C terminus of each $\alpha 2$ helix extends further along the long axis than the adjacent N terminus of the paired histone (Figs 1d, 2a, b). This effectively tapers the ends of the pair to create a similar type of DNA-binding site, here called L1L2, at each end of the dimer. Within the histone fold, the N-terminal $\alpha 1$ helix and C-terminal $\alpha 3$ helix fold back on roughly the same side of the central $\alpha 2$ helix, each generating a left-handed spiral in the intervening loops. This arrangement gives a single histone a broad, skewed U-shaped appearance, with two Us slotting together to form a pair. The short, 3-turn $\alpha 1$ and $\alpha 3$ helices make contact with both $\alpha 2$ helices in the dimer, but because of the offset in the $\alpha 2$ to $\alpha 2$ packing, only the $\alpha 1$ helices and their short extensions make contact between opposite histones. The $\alpha 1$ surfaces complete the convex DNA-binding surface, providing its centre and second type, or $\alpha 1\alpha 1$, of DNA-binding site. On the opposite, concave side of the dimer, the $\alpha 3$ helices are not in contact.

The formation of only H3-H4 and H2A-H2B heterodimeric pairs instead of homodimers or other heterodimers results from complementary internal packing, as seen from alignment of the structures, although the packing is complicated by further requirements for histone pair-pair and histone-DNA interactions. Obvious differences do occur in the manner in which the $\alpha 1$ helices, their N-terminal extensions, and the L1, L2 loops interact within a pair. Alignment of the four histone types based on the $\alpha 2$ helices reveals that the L1 loop lengths for H3 and H2B are 8 amino acids, whereas for H4 and H2A they are 6 and 7 amino acids, respectively

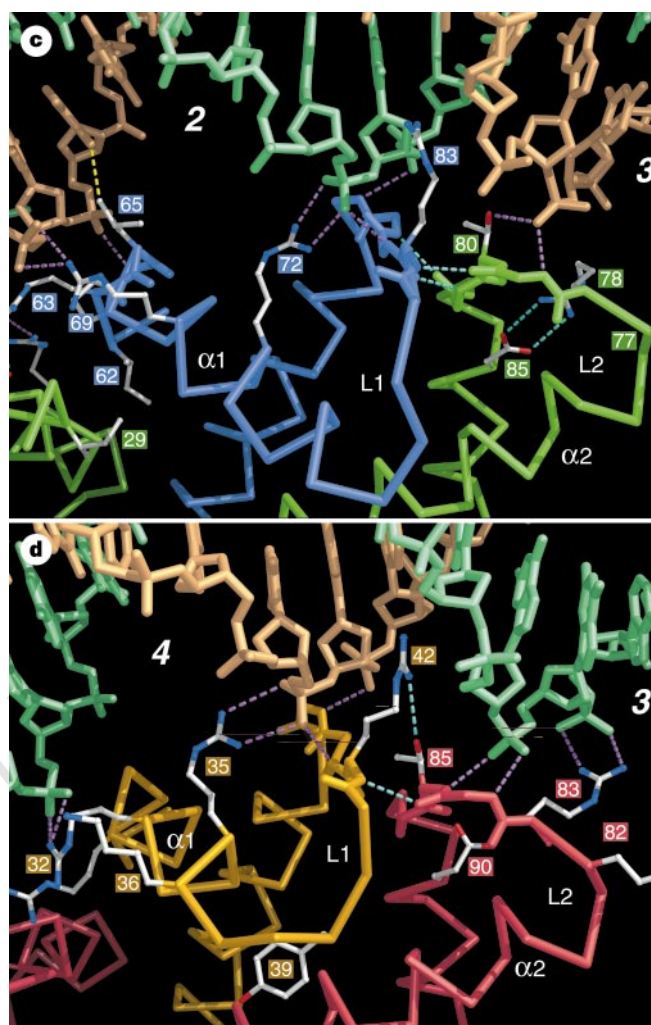
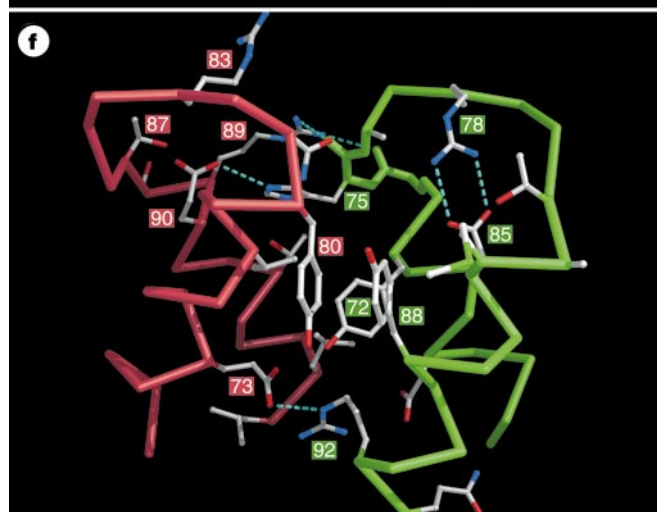
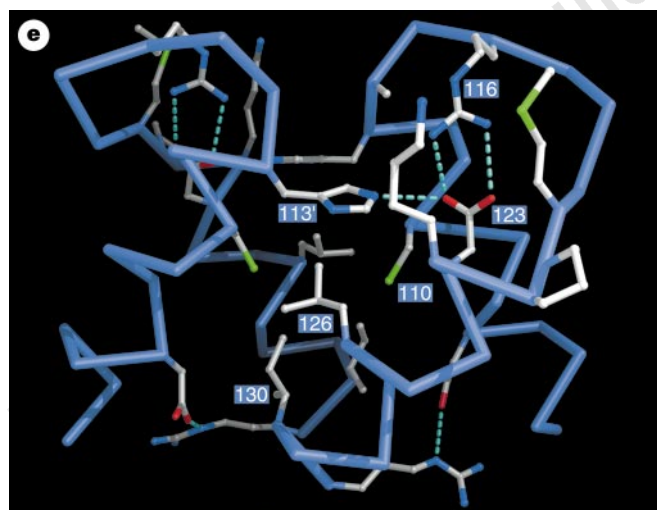
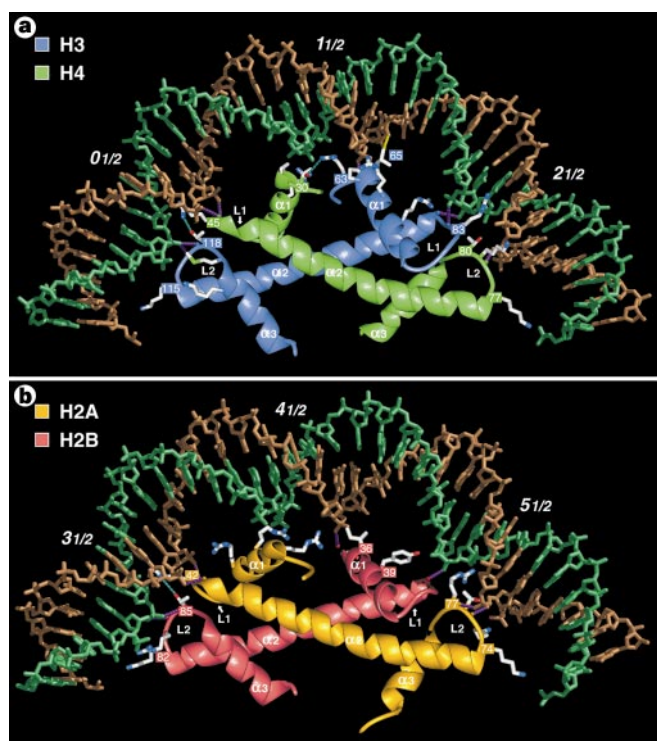


Figure 2 **a**, H3-H4 histone-fold pair. The α 1-L1- α 2-L2- α 3 structural elements are shown. A pseudodyad axis of symmetry runs vertically through SHL1.5. Side chains that make hydrogen bonds or hydrophobic interactions with the DNA backbones, and arginines inserted in the minor groove are shown. Main-chain-to-DNA-phosphate hydrogen bonds are shown in magenta. Leucine 65 is in contact with a thymidine methyl group (yellow bond). Two of three arginine-threonine pairs are hydrogen bonded (cyan bond). **b**, H2A-H2B histone-fold pair. As in **a**, except that the pseudodyad axis runs through SHL4.5. **c**, H3-H4 L1L2 DNA-binding site. The H3 L1 and H4 L2 loops make 3 hydrogen bonds (cyan) with each other in a parallel β -structure. The L2 loop contains buried hydrogen bonds (cyan) between H4-R78 and H4-D85. The hydrogen-bonding interactions (magenta) between protein and DNA-phosphate groups involve both main-chain amide atoms and side-chain atoms. **d**, H2A-H2B L1L2 DNA-binding site. In contrast to the H3 L1 and H4 L2 loops, the other three L1L2 sites have only one well formed β -structure hydrogen bond, as for H2A L1 and H2B L2 shown here (cyan). An arginine (H2B-R83) hydrogen bonds to a phosphate group instead of an acidic side chain (H2B-E90), as occurs at the homologous sites in the other L1L2 loops. **e**, H3'-H3 4-helix bundle. The H3-H4 histone pairs form the tetramer through the interaction of the C-terminal halves of the α 2 helices and the α 3 helix of H3' and H3 across the molecular dyad axis. Histidine 113' makes a buried hydrogen bond to link the H3 molecules. In addition, several hydrophobic interactions occur. **f**, H2B-H4 4-helix bundle. The H2A-H2B dimer assembles into the histone octamer through the interaction of H2B with H4. The hydrogen bond of H4-H75 to H2B-E90 is analogous to that in H3'-H3, but is not buried here because of the orientation of H2B-R83. Across the pseudo-twofold axis, however, H2B-Y80 replaces the H4 histidine and, together with H4-Y72 and H4-Y88, forms a hydrophobic cluster.

(Fig. 1c). The loops run essentially antiparallel to the connected $\alpha 1$ helices, so that the shorter loops of H4 and H2A are compensated for by an earlier stop of the $\alpha 1$ helix with respect to the underlying contacts. The side chains I62 of H3 and I29 of H4 on the $\alpha 1$ helices of a H3–H4 pair pack together with pseudodyad symmetry (Fig. 2c), but the H2A–H2B dimer has an asymmetric interface here because of the insertion of Y39 of H2A into the H2A L1 loop, as compared to L1 for H4 (Figs 1c, 2d). Outside the histone folds, the αN extension of H3 and the αC extension of H2B contribute substantially to their respective pairings.

The extra length of the L1 loops and extended N terminus of the $\alpha 1$ helices for H3 and H2B take a path somewhat further from the overall long axis of a histone pair compared to the L1 loops of H4 and H2A. The adjacent H4 L2 and H2A L2 loops are thereby displaced towards this central axis so that the C termini of the H4 and H2A $\alpha 2$ helices are bent towards the L1 loops of the partner: the H4 and H2A $\alpha 2$ helices curve, whereas those of H3 and H2B are straight.

The L2 loops in the C-terminal half of the histone fold show a high degree of structural homology between histones. A three-amino-acid stretch in the L2 loop of one histone in a pair and the L1 loop of the other run roughly parallel to each other at each end of a dimer. In the case of H3 L1 with H4 L2, the maximum of three hydrogen bonds are made in parallel β -conformation (Fig. 2c), whereas for the other three L1–L2 interactions, only one hydrogen bond is evident. The conformation of these loops may depend on the DNA sequence bound, which can be tested by obtaining structural data from particles with different DNA fragments. A second important feature of the H3 L2 and H4 L2 loops is that although arginines H3-R116 and H4-R78 are in close proximity to the DNA backbone, they face away from it and make hydrogen bonds, respectively, with aspartate side chains H3-D123 and H4-D85 (first turn of the $\alpha 3$ helices). These acidic side chains are inaccessible to solvent contact. In contrast, the equivalent H2B-R83 does bind a DNA phosphate, despite the presence of H2B-E90 at the same position as the aspartate groups in H3 and H4 (Fig. 2d), whereas H2A lacks both the basic and acidic side chains at these sites.

The stable oligomeric aggregates of histones are the H3–H4 tetramer and the H2A–H2B dimer in solutions of physiologically relevant ionic strength, or the histone octamer in the presence of DNA or high salt concentration. The C-terminal halves of the $\alpha 2$ helices and the $\alpha 3$ helices are the primary determinants of the tetramer and octamer assemblies through the formation of three four-helix bundles—a single H3'–H3 association to form the tetramer (Fig. 2e) and two H2B–H4 associations to produce the octamer (Fig. 2f). Including the histone folds and extensions, further interactions occur within an octamer, between H3–H2A', H4–H2A', H4–H2B', H2A–H2A' and H2A–H2B'. The use of the H3, H4 and H2B $\alpha 3$ helices as the principal interaction sites for the assembly of tetramer and octamer suggests that the remaining H2A $\alpha 3$ helix might form an internucleosomal 4-helix bundle to create the higher-order structure of nucleosomes. However, the resulting complex seems unlikely to be formed, based on current models for the higher-order structure³.

The H3'–H3 4-helix bundle contains buried charged groups which are hydrogen-bonded in the intermolecular interaction (Fig. 2e). These interactions involve H3-D123 again and, in addition, H3'-H113 on the adjacent copy of H3. The side chain of cysteine H3-C110, which has been the primary target for attachment of extrinsic probes of nucleosome structure, is in the centre of the bundle. Access to this site would require the H3-L126 side chain to adopt an alternative conformation, which is sterically possible. Core particles prepared in oxidizing conditions have the H3-C110 side chains linked in a disulphide, and these particles have somewhat different properties from those kept in fully reducing conditions²⁴. The 6.2 Å separation of the sulphhydryl groups is too

Table 1 Crystallographic analysis

Diffraction data	Native	1 × TAMM	2 × TAMM	4 × CH ₃ HgNO ₃
Resolution	25–2.8 Å	25–2.8 Å	25–2.8 Å	25–2.8 Å
Observations	297, 397	235, 751	229, 513	186, 451
Unique <i>hkl</i>	52, 432	52, 345	52, 372	50, 986
Completeness (%)	0.99	0.98	0.98	0.96
<i>R</i> _{merge} [†]	0.056	0.064	0.068	0.056
<i>R</i> _{merge} [‡] 3.0–2.8 Å	0.132	0.135	0.127	0.129
MIR analysis				
Heavy-atom sites		H3 C110	H4 T73	H3 E133, H4 S47
Phasing power [§] (MLPHARE)		0.84	0.92	0.97
Cullis <i>R</i> -factor [¶]		0.89	0.89	0.85
Refinement statistics				
Observations (<i>F</i> > 3 × σ), working/test			43,638/2,639	
<i>R</i> -factor [§] /free <i>R</i> -factor (%)			22.4/30.2	
Component	Number of atoms	R.m.s.d. bond length	R.m.s.d. bond angle	Mean <i>B</i> -factor
Protein	6,387	0.010 Å	1.92°	40.4
DNA	5,980	0.006 Å	1.13°	80.2
Total	12,367	0.009 Å	1.31°	59.7

^{*} $\sum_{hkl} |I_{hkl} - \bar{I}_{hkl}| / \sum_{hkl} I_{hkl}$, where $\bar{I}_{hkl}$ is the mean of measurements for a single *hkl*.

[†] $\left\{ \sum_{hkl} F_{hkl}^2 / \sum_{hkl} \{F_{PH,obs} - F_{PH,calc}\}^2 \right\}^{1/2}$.

[‡] $\sum_{hkl} |F_{PH} \pm F_P| - F_{H,calc} / \sum_{hkl} |F_{PH} \pm F_P|$, centric only.

[§] $\sum_{hkl} |F_{obs} - F_{calc}| / \sum_{hkl} F_{obs}$.

large for disulphide bond formation without substantial distortion to the tetramer conformation. The H3-C110 amino acids were the target of the tetrakis(acetoxymethylmercurimethane) (TAMM) multi-heavy-atom compound that we used to produce single-site derivative crystals for the structure determination (the H3-C110 sulphhydryls are bound and the H3-H113 imidazoles are not). For other derivatives, the H3-C110A mutant was used to avoid disturbance of the structure and to avoid duplicate sites between derivatives.

The H2B–H4 4-helix bundle is closely related in its overall conformation to that for H3'–H3 (r.m.s.d. on α -carbon of 1.85 Å). Histidine H4-H75 makes a buried hydrogen bond with glutamate H2B-E90, but the twofold related interaction is completely different (Fig. 2e, f). A tyrosine side chain, H2A-Y72, replaces H3-C110 and lies flat against the tyrosine side chain H2B-Y80, where one of the buried histidine residues is found in H3. Both H3'–H3 and H2B–H4 bundles contain several other hydrophobic and hydrogen bonds. The increase in hydrophobic contribution to this H2B–H4 interface compared to that in H3'–H3 explains, at least in part, the instability of the histone octamer in contrast to the tetramer at low salt concentration.

The remaining interactions between the histone-fold domains occur between H2A, H2B and H2A'. Glutamate H2A-E41 and histidine H2B-H79 are organized each to make a hydrogen bond with the side chain of glutamine H2A'-N38. A second pair of identical bonds is made nearby across the particle pseudo-twofold axis. These interactions suggest that there is cooperativity in dimer binding.

Histone-fold/DNA interactions

The histone-fold domains account directly for the organization of 121 bp of DNA. Each histone-fold pair, whether half of the H3–H4 tetramer or a H2A–H2B dimer, is associated with 27–28 bp of DNA, leaving 4-bp linkers between these units (Fig. 1d, 2a, b). Binding is primarily to the DNA phosphodiester backbones as they face the protein over 2.5 turns of the double helix. Each helical turn of a chain commits a segment of only two consecutive phosphate groups to hydrogen-bonding interactions with the histone-fold pairs. The length of the DNA bound is extended by an additional backbone segment in both directions, eight in total (34–36 bp), if

the H3-K115, H4-K77, H2A-K74 and H2B-K82 conserved lysines that begin the L2 loops and make salt links to phosphate groups are included. The extended DNA sites overlap from one histone-fold pair to the next.

The histone-fold DNA-binding sites can be divided into two types. The first, or $\alpha 1\alpha 1$ site, uses both $\alpha 1$ helices to bind the two DNA backbone segments at the centre of the bound DNA stretch. The second, or L1L2, sites are formed from L1 and L2 loops and termini of the $\alpha 2$ helices at each end of the histone-pair crescent. Each binds to the two DNA backbone segments at the edge of the whole DNA site. In general, five predominant features of the histone-fold/DNA interaction are observed each time the phosphodiester chain faces the histone octamer (Fig. 2a–d). (1) The N termini of the $\alpha 1$ helices of H3, H4 and H2B, and the $\alpha 2$ -helices of all four core histones, are each used to fix the position of an individual phosphate group through the positive charge generated by the helix dipole. (2) Hydrogen bonds to phosphates are made from main-chain amide nitrogen atoms of amino acids either near or in the last turn of the $\alpha 1$ and $\alpha 2$ helices. (3) An arginine side chain from a histone fold enters the minor groove 10 of the 14 times it faces the histone octamer. The other four occurrences have arginine side chains from tail regions penetrating the minor groove. (4) Extensive nonpolar contacts are made with the deoxyribose groups. (5) Hydrogen bonds and salt links occur frequently between DNA-phosphate oxygen atoms and protein basic and hydroxyl side chain groups. These side chains approach the inward-facing DNA backbone segments from both the major and minor groove sides.

In the $\alpha 1\alpha 1$ DNA-binding site, H3, H4 and H2B all have the N terminus of their $\alpha 1$ helix pointing at an individual phosphate group. For H3 and H4 at SHL1.5, proline side chains in the first turn of $\alpha 1$ contact deoxyribose groups (H3-P66, H4-P32) and H3 main-chain amide nitrogen atoms (H3-K64, H3-L65) hydrogen-bond to phosphate oxygen atoms. In the case of H2B at SHL4.5, the $\alpha 1$ helix begins one turn later in the structure superposition of the histone folds (Fig. 1c), and isoleucine H2B-I36 now contacts a deoxyribose while the main-chain amide nitrogen of H2B-S33 hydrogen-bonds to phosphate. As a consequence of the orientation of the H3 and H4 $\alpha 1$ helices relative to those of H2A and H2B, the DNA is 3.5 Å displaced from the underlying $\alpha 2$ helices of a H3–H4 pair compared to a H2A–H2B pair. On the side of the $\alpha 1$ helix that faces the DNA major groove, the side chains (of arginine, lysine, histidine and tyrosine) generally hydrogen-bond to phosphate groups (H3-R69, H4-R35, H2A-R32) or create a complementary, positive electrostatic charge density (Fig. 2a, b). In one case, H3 leucine 65 reaches into the major groove and makes a hydrophobic interaction with the 5-methyl group of thymidine (Fig. 2c). The minor groove at the centre of the binding site for the H3–H4 pair has H3-R63 inserted between phosphate chains (Fig. 2a), whereas the structurally analogous site on H4 contains a threonine group in contact with deoxyribose. These threonine and arginine side chains hydrogen-bond, restraining the arginine guanidinium group from penetrating more deeply into the DNA groove. The H2A $\alpha 1$ helix is oriented differently with respect to the DNA because of the insertion of Tyr 39 in the L1 loop. In this case, the phosphodiester chain is

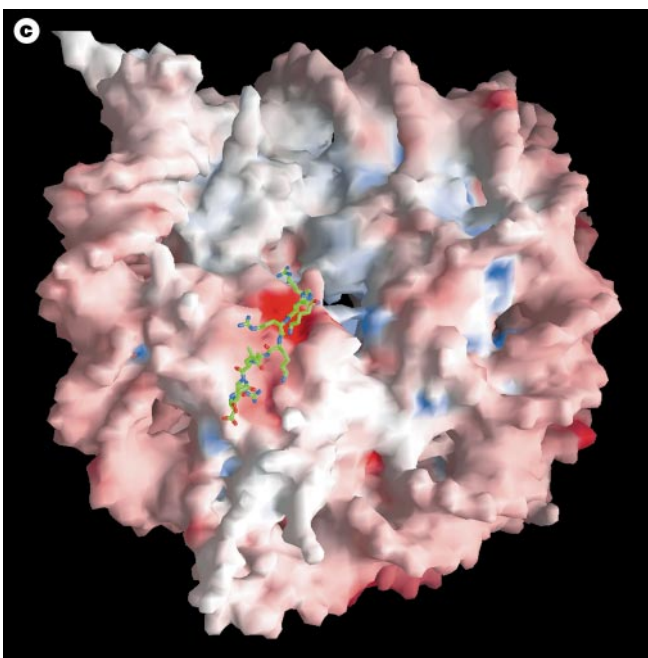
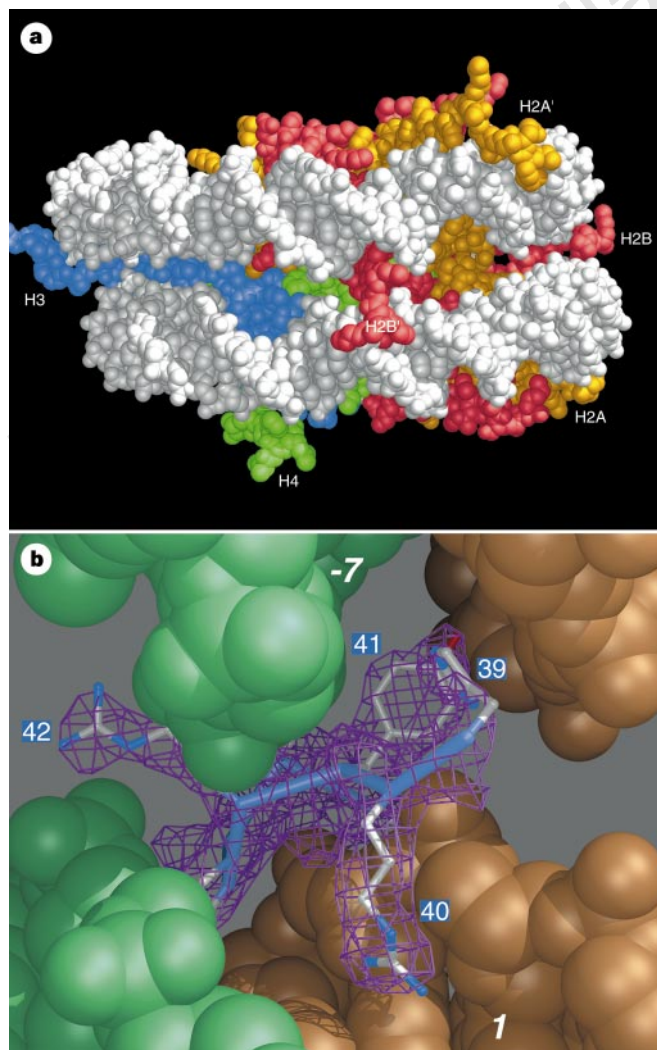


Figure 3 a, Histone tails between DNA gyres. The H2B (red) and H3 (blue) N-terminal tails pass through channels in the DNA superhelix (white) formed by aligned minor grooves. **b**, H3 tail segment in minor-groove channel. The H3 tail sequence HRYRP passes through the juxtaposed minor grooves of SHL –7 (DNA terminus) and SHL1 (DNA centre). The electron density (magenta) clearly identifies the conformation of these amino acids. **c**, H4 tail bound to the H2A–H2B dimer. The electrostatic surface of the nucleosome core particle shows an intensely negatively charged region (bright red) near the centre of its exposed protein face. Amino acids 16–25 of the H4 tail lie on this region of the H2A–H2B dimer, which contains 7 clustered acidic side chains. The orientation of the particle is as for Fig. 1d.

held by H2A arginines (R29, R32) extending from the $\alpha 1$ helix. The N termini of the $\alpha 1$ helices of the H2A–H2B dimers are the takeoff points for N-terminal tails that interact with the DNA. The tail of H2A clasps the DNA at the minor groove on the outside of the particle, and the tail of H2B passes through the minor groove channel between superhelix gyres. At SHL4.5, the guanidinium group of H2B-R30 in the tail penetrates most deeply into the minor groove.

In an L1L2 DNA-binding site, the N terminus of an $\alpha 2$ helix points towards a phosphate group and, for H3, H4 and H2A, the arginine side chain in the preceding L1 loop is inserted into the minor groove (Fig. 2c, d). The main-chain amide nitrogen atoms in the L1– $\alpha 2$ junction hydrogen-bond to the DNA phosphates, probably specifying their positions precisely. The H4 and H2A–L1 loops are nearest to the pseudodyad axis of the core particle at SHL0.5 and 3.5, respectively, whereas the longer H3 and H2B–L1 loops are more distant at SHL2.5 and 5.5, respectively. The more highly conserved L2 loops at the $\alpha 2$ C termini bind the outermost DNA strands of the bound 26–28-bp DNA segment and hold the phosphates in place with hydrogen bonds from main-chain amide nitrogen atoms as well as side chains. The single arginine in the L1 loop of each H3, H4 and H2A (H3-R83, H4-R45, H2A-R42), which extends into the DNA minor groove in each case, is hydrogen-bonded to a threonine hydroxyl group in the adjacent L2 loops of H4, H3 and H2B (H4-T80, H3-T118, H2B-T85), respectively. This interaction appears to restrain the arginine side-chain conformation so that it is not free to hydrogen-bond with the nearest O2 atom of cytosine or thymidine, or the N3 of adenine or guanine. In one case, however, H4-R45 hydrogen-bonds to the O2 of a thymidine base at SHL0.5 while maintaining the threonine interaction. In the H2B sequence, the L1-loop arginine is replaced by glycine (H2B-G50) and the position of the restraining threonine is replaced by H2A-R77, which penetrates the DNA minor groove at SHL5.5. As the DNA straightens at this location, probably because there are fewer histone contacts compared with other binding sites, H2A-R77 does not extend further into the minor groove than the arginine–threonine pairs.

Tails and extensions

The N-terminal tails of both H3 and H2B have random-coil segments that pass between the gyres of the DNA superhelix (Figs 1c, 3a). For H3, five amino acids (39–43) pass through a channel in the superhelix formed by the minor grooves of the DNA terminus (SHL6.5 to 7) and central turn (SHL –0.5 to –1) (Fig. 3b). An extremely basic, 8-amino-acid span of H2B also passes through the superhelix (SHL4.5 and SHL –2.5 to –3), which creates, with H3 and the twofold molecular symmetry, a 20-bp periodicity of tails in minor-groove channels. At a location (SHL4 to 4.5) adjacent to the H2B interaction, four amino acids of the H2A N-terminal tail are bound to the minor groove on the outside of the superhelix (Fig. 3a). The electron density for tail sequences further N-terminal is weak in most cases and therefore not interpreted. These regions, which are highly basic and contain the sites of acetylation for H3 and H2B and border the ubiquitination sites for H2A, are probably involved in formation of the chromatin fibre through interactions with neighbouring nucleosomes.

The H3 αN helix extension of the histone-fold and preceding tail region in the minor grooves are responsible for binding the 13 bp of DNA at each terminus of the superhelix (Fig. 1b, d). The αN helix lies on the H4 $\alpha 2$ N terminus and H4 L1 loop on the opposite face of the particle from the $\alpha 1$ helix to which it is connected. A section of the H2A' C-terminal tail (amino acids 105–117) runs antiparallel to αN and further links it to the underlying H3–H4 histone-fold domains.

Preceding and overlapping the region of the H2A C terminus that contacts H3' αN , the H2A tail (amino acids 92–108) forms a folded 'docking' domain with the H2A $\alpha 3$ helix. This buried domain contains a short α -helix (amino acids 92–96), and makes β -sheet

interactions with the short H4' C-terminal tail (amino acids 95–102) folded back over its $\alpha 3$ helix (Fig. 1d). The H2A tail and docking domain (amino acids 80–119) account for a loss of 1,914 Å² in accessible surface area on binding of a H2A–H2B dimer to the H3–H4 tetramer, as compared to 1,074 Å² for the H2B–H4 4-helix bundle interaction. These extensions of the histone-fold motif appear to be necessary to stabilize the helical ramp formed by the histone-fold domains, maintaining its pitch accurately.

A striking feature of the protein surface is the V-shaped pair of antiparallel H2B αC and $\alpha 3$ helices (Fig. 1d; side chains not shown). The αC helix defines the outer limit of the flat face of the core particle and is held in place by a 4-amino-acid linker to the preceding H2B $\alpha 3$ helix and by extensive hydrophobic interaction with the H2A $\alpha 2$ helix, including H2A-Y50, which is inserted between the two H2B α -helices. The tyrosine H2B-Y118 at the C terminus of the αC helix interacts with a single-turn α -helix within the H2A N terminus (amino acids 17–20) before this tail passes over the DNA minor groove at SHL4 to 4.5 (Fig. 1d). The N-terminal tail of H2B passes through the minor groove channel just preceding SLH –3 and +5, with only H2B-E32 linking this DNA-binding region to the histone-fold domain. The requirement of immediate entry of the H2B tail into the adjacent minor groove is accommodated by the limited interaction between the H2B and H2A $\alpha 1$ helices.

The two H4 N-terminal tails up to amino acid 25 have divergent structures. Only one is well localized and then only in the stretch from K20 to I26, although diffuse electron density exists additionally for K16 to R19. The observed part makes extensive contact with an H2A–H2B dimer of an adjacent particle. This H4 region, K16 to N25, makes multiple hydrogen bonds and salt bridges between its basic side chains (K16, R19, K20, R23) and acidic side chains of H2A (E56, E61, E64, D90, E91, E92) and H2B (E110). Because of the positive charge of the H4 tail and negative charge on the H2A–H2B dimer surface (Fig. 3c), it is uncertain whether the observed interactions are unique. The importance to successful crystallization of one of these interactions was confirmed as the mutant with K20C cannot crystallize under our conditions, but can be overcome through reintroduction of the ϵ -amino group by making the H4-C20–cystamine adduct (K.L. and T.J.R., unpublished results).

DNA superhelix

The diffraction quality of nucleosome core particle crystals depends largely on the DNA sequence that is incorporated. Through trials of many sequences, we have found that two different halves of human α -satellite repeats yield diffraction to 2.0 Å resolution when used as an inverted repeat. The potential for structure determination of a mutant of one of the α -satellite repeats has been noted²⁵. We were able rapidly to check the position(s) of octamer on a variety of different DNA sequences by using our site-specific, hydroxyl-radical cleavage method⁷. The X-ray structure verifies our discovery made by using this technique, namely that the molecular two-fold axis of the particle passes through a single base pair and not between two, despite the use of a 146-bp palindromic sequence (Figs 1b, 4a). We can identify directly from the electron density the type of base pair for ~50% of the sequence (Fig. 4b, c), and thereby distinguish the 72-bp from the 73-bp half extending from the central base pair. Termini of the 72- and 73-bp halves from adjacent particles pack end-to-end in a pseudo-continuous double helix, but with a twist of 132° at the junction (Fig. 4d). Superposition of the two DNA halves shows that they have closely similar structures apart from a 12-bp section of the 72-bp half (base pairs 12–23) which covers a 13-bp region in the 73-bp half (Fig. 4d). We attribute the location of this stretched region either to strong, specific DNA-positioning requirements in the flanking sequences, or to the strength of the inter-particle contact between DNA termini. The latter explanation is favoured because a second 146-bp DNA sequence that yields

high-quality crystals displays a similar distortion, but it occurs only two turns away from a DNA terminus (data not shown).

The best ideal superhelix fit to the double helix axis over base pairs from 62 in the 72-bp half to 63 in the 73-bp half has a 41.8 Å radius and a 23.9 Å pitch. As the 10-bp segment at each terminus is essentially straight, the effective number of superhelical turns is 1.65. The DNA locations most distant from the superhelix axis, SHL ± 1.5 (44 Å) and ± 4 to 5 (45 Å), also show the greatest bending. The underlying histone structures are the $\alpha 1\alpha 1$ DNA-binding sites of the H3–H4 pairs for SHL ± 1.5 (Fig. 2a), which cause an outward bulge in the DNA, and the adjacent L1L2 and $\alpha 1\alpha 1$ DNA-binding sites of the H2A–H2B dimers, which buckle the DNA outwards between them. In general, the superhelix path is significantly distorted owing to the local structure of the histone DNA-binding surface.

The parameters of the double helix also show substantial variation over the length of the superhelix. The overall twist, for example, is 10.2 bp per turn in the local reference frame (independent of superhelical pitch). Dividing the DNA into segments delimited by the arginine side chains inserted into the minor groove, the minimum value is obtained for the overwound stretch in the 72-bp half that compensates for one less base pair: 9.4 bp per turn versus 10.6 for the same segment in the 73-bp half. The maximum value of 10.9 for unwinding occurs at the SHL-5 segment, where an H2B N-terminal tail passes through the minor-groove channel. When applied to the left-handed superhelix, the overall overwinding of nucleosomal DNA by 0.3 bp per turn compared to free DNA creates an alignment of minor grooves between superhelix gyres that allows an essentially straight passage for the H3 and H2B tails through the channels formed. Further details of the nucleosome DNA parameters

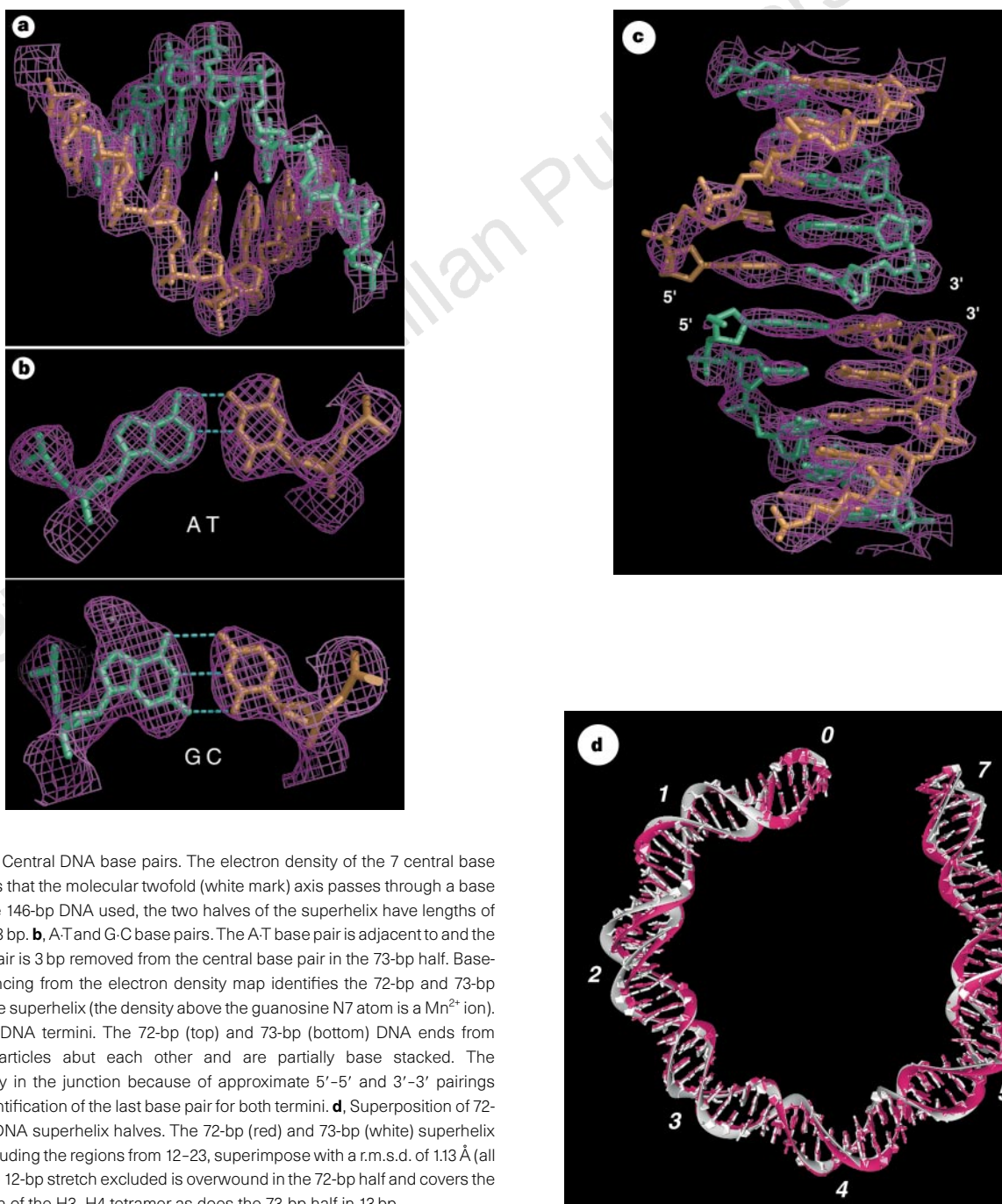


Figure 4 **a**, Central DNA base pairs. The electron density of the 7 central base pairs shows that the molecular twofold (white mark) axis passes through a base pair. For the 146-bp DNA used, the two halves of the superhelix have lengths of 72 bp and 73 bp. **b**, A-T and G-C base pairs. The A-T base pair is adjacent to and the G-C base pair is 3 bp removed from the central base pair in the 73-bp half. Base-pair sequencing from the electron density map identifies the 72-bp and 73-bp halves of the superhelix (the density above the guanosine N7 atom is a Mn^{2+} ion). **c**, Stacked DNA termini. The 72-bp (top) and 73-bp (bottom) DNA ends from adjacent particles abut each other and are partially base stacked. The discontinuity in the junction because of approximate 5'–5' and 3'–3' pairings permits identification of the last base pair for both termini. **d**, Superposition of 72- and 73-bp DNA superhelix halves. The 72-bp (red) and 73-bp (white) superhelix halves, excluding the regions from 12–23, superimpose with a r.m.s.d. of 1.13 Å (all atoms). The 12-bp stretch excluded is overwound in the 72-bp half and covers the same region of the H3–H4 tetramer as does the 73-bp half in 13 bp.

will be reported in a comparison of two nucleosome structures containing different DNA sequences (A.M., K.L. and T.J.R., manuscript in preparation).

Discussion

The significantly non-uniform shape of the DNA superhelix and semiregular variation in DNA parameters suggests that nucleosome positioning results in part by DNA-sequence-dependent flexibility and histone-induced bendability. The double-helix twist value of 10.2 bp per turn observed for the structure overall matches the sequence periodicity of DNA taken from nucleosome cores²⁶ but deviates from the 10.5-bp-per-turn twist value measured for free DNA²⁷, indicating that eukaryotic DNA has adapted to fit the nucleosome. Conversely, DNA sequence should affect nucleosome formation and stability. The observation that G-C base pairs are preferred when the major groove faces inwards is simply explained by their tendency to bend into the major groove²⁸. Furthermore, the single arginine side chains that are inserted into the minor groove at every turn of the double helix impart a sequence preference to those base pairs in contact with the histones. Sequences that can most easily adopt a narrow minor groove (correlated with bending into the minor groove and a local helix overwinding, as observed in the structure) will be preferred because stronger salt links can be made between phosphates across the groove, resulting in the preference for A-T base pairs in these locations²⁹. In addition to sequence-dependent properties that can affect nucleosome positioning, specific and semispecific interactions can occur, such as that between a leucine residue and a thymidine 5-methyl group or an arginine hydrogen-bonding to a pyrimidine O2 atom. The generation²⁶ and characterization⁷ of a nucleosomal DNA sequence database, when analysed in terms of the actual nucleosome core particle DNA structural template, will enable rules to be drawn up for predicting nucleosome position.

The packaging of DNA in nucleosomes generally inhibits the binding of non-histone, DNA-binding proteins such as transcription factors. The restrictions imposed by bound histones may be as much due to distortion of the DNA as to the direct blocking of binding sites. The static accessibility of the DNA molecule bound in the core particle to a probe of 4 Å diameter is 73% of the total value for the bare superhelix. In principle, a well positioned nucleosome could provide factor-binding sites that are distinct from any available in free DNA. The HIV integrase, for example, preferentially inserts the HIV genome into nucleosomes with two apparent hot spots, where the major groove faces outwards at SHL 1.5 and SHL 3.5 to 4 (refs 30, 31). These two sites have the highest curvature of locations in which the major groove faces outwards. The histone tail segments passing through the minor-groove channels every 20 bp may reinforce the periodicity of the integration pattern.

Most DNA-binding factors need DNA to dissociate from the nucleosome in the region of their binding site to allow access. Because of the multiple DNA-binding sites of the octamer (9–10 pairs of chemically distinct structures: four L1L2, two α 1 α 1, 3–4 tails), dissociation would occur in semi-cooperative stages (non-contacted base pairs in one stretch are maximally four), beginning at the entry and exit points of the DNA held by the H3 N-terminal tail and α N extension, and then proceed into the H2A–H2B dimer, and eventually into the H3–H4 tetramer regions. This scheme was confirmed by measurement of the accessibility of restriction enzyme sites in nucleosome cores, which showed an increasing resistance to digestion of the order of 10^2 for sites at DNA termini, to 10^5 for sites at the DNA centre, compared to free DNA³². When a DNA-binding factor approaches nucleosomal DNA, the reduction of cooperativity in binding to the histones appears as cooperativity in the binding of successive protein factors, even without their direct interaction¹³. The affinity of one to a few histone sites for DNA is not insignificant, however, even in the region of SHL6–7, because transcription by a bacteriophage RNA polymerase is impeded first when it encounters

a nucleosome and then again when it reaches SHL5, the location just before the site of interaction with the H2B N-terminal tail. Apparently the polymerase pauses when the DNA behind it binds the histone octamer at SHL6–7 and creates a constraining loop³³. For many genes, however, ATP-dependent nucleosome-modification complexes, such as the yeast SWI/SNF complex, are required to permit transcription in the presence of nucleosomes³⁴. Remodelling factors may simply aid the dissociation of DNA from the histone octamer, for example, by destabilizing the H2A–H2B dimer/H3–H4 tetramer interactions. Mutations in yeast histones H3 and H4 that partially overcome the loss of the SWI/SNF complex are found predominantly in the H4 L1 loop and the adjoining H3 L2 loop at SHL0.5 (ref. 35). The specific residues affected are the buried H3–R116, H4–R45 inserted in the minor groove, and H3–T118, which is hydrogen-bonded to both H4–R45 and a DNA phosphate group. Additional residues with the same phenotype are H4–V43, which contributes to the shape of the loops by contacting an analogous H3 isoleucine, and H3 E105, for which a direct effect is difficult to envisage at the mononucleosome level.

A model for the higher-order structure of chromatin can be built using our structure of the nucleosome core particle that is highly consistent with the 'standard' model of Finch and Klug^{36–39} (T.J.R., unpublished observation). A remarkable feature suggested by this model is that the highly basic, H4 N-terminal tail sequence from amino acids 1–25 binds as an extended chain to a region of extreme acidity on the exposed face of the H2A–H2B dimer (Fig. 3c). This region clearly binds the H4 amino acids 16–24 in the interparticle contacts of the crystals. Although in our model of higher-order structure, the H4 tail is incorporated in the opposite directional sense compared to that in the crystal, the acidic surface and basic tail should permit an equally extensive interaction in either orientation. In the crystal, the H2A–H2B dimer binding site for the H4 tail includes seven clustered acidic amino acids. The H4 amino-acid stretch 16–24, which is bound between particles, binds to the yeast silencing protein SIR3 and participates in the formation of yeast 'heterochromatin'⁴⁰. Our model suggests that an internucleosome connection in the higher-order structure is replaced by an interaction with SIR3. In the model, the four lysine acetylation sites at the H4 N terminus are near the H2B α C-helix. The affinity of the crossbridge formed by the H4 tail to the open face of the adjacent nucleosome in the model of the higher-order helix would be reduced by successive acetylation of the H4 lysines 5, 8, 12 and 16. The role of acetylation in releasing DNA bound in chromatin is more likely to be a destabilization of the nucleosome higher-order helix than of the nucleosome itself.

The histone-fold motif is highly conserved, as seen from structures obtained from organisms as diverse as archaeal bacteria⁴¹, insects⁴², birds²⁰ and amphibians (this study), presumably because of its unique DNA-binding properties. Nevertheless, the core histones occur in four types which may have diverged to impart particular characteristics to different segments of the DNA superhelix, and to accommodate requirements of the higher-order structure and interaction with assembly, transcription and remodelling factors. Histone sequences are more variable outside the histone fold; these extensions and tail regions may be absent, as in the case of the archaeal proteins⁴¹, or replaced by entirely different domains, as in the *Drosophila* factors TAF_{II}42 and TAF_{II}62 and their homologues⁴². For these TAFs and TAF30 α , which is the most similar to H2B, the arginine side chains of the core histones that insert into the minor groove are replaced by other side chains, suggesting that if these TAFs do bind DNA, then the mode of interaction is likely to be different from that in the nucleosome. □

Methods

Crystal preparation. Nucleosome core particles were prepared from a 146-bp palindromic DNA fragment derived from human α -satellite DNA⁴³ and recombinant histone proteins (Fig. 1b, c)²³. Crystals were grown by vapour

diffusion in 8–20 days at 20 °C using a droplet containing 4 mg ml⁻¹ core particle, 50 mM KCl, 70–75 mM MnCl₂, and 20 mM potassium cacodylate, pH 6.0, surrounded by silicon oil DC200 (110 mPa s; Fluka) and equilibrated against 40–46 mM MnCl₂, 35–40 mM KCl and 20 mM potassium cacodylate, pH 6.0. Crystals were transferred to 37 mM MnCl₂, 40 mM KCl, 20 mM cacodylate, pH 6.0, and 24% 2-methyl-2,4-pentanediol for stabilization and improvement of diffraction resolution. Three heavy-atom derivatives, selected from 11 crystal types each with a different mutant particle, were made by soaking crystals with TAMM⁴⁴ (1 × TAMM), or by solution-labelling core particles containing the mutants H4-T73C with TAMM (2 × TAMM), or H3-E133C and H4-S47C (4 × CH₃HgNO₃) in 20 mM Tris-acetate, pH 7.5, with methyl-mercury nitrate before crystallization.

Crystallographic analysis. Data are summarized in Table 1. Crystals belong to space group *P*2₁2₁2₁, with one nucleosome core particle per asymmetric unit and unit cell parameters: *a* = 106.04, *b* = 181.78, *c* = 110.12. For data collection, crystals were mounted in tapered glass capillaries, transferred to 4 °C, cooled over 5 min to -18 °C, flash-cooled in liquid propane at -120 °C, and transferred into a N₂ gas stream at -170 ± 5 °C. X-ray intensities were collected using a 30-cm MAR image plate scanner at a wavelength of 0.97 Å (ID13, ESRF, Grenoble). Data (25–2.8 Å) were processed using local programs (T.J.R., unpublished) and heavy-atom sites were identified from difference Patterson maps. MIR phases and model combination were calculated using local programs (T.J.R., unpublished) and yielded an overall figure of merit of 73.5 (46.7 with MLPHARE).

Model building and refinement. The initial MIR map was used to trace 120 bp of DNA and 745 amino acids. Iterative rounds of model building with O⁴², simulated annealing refinement with XPLOR⁴⁶, and phase combination with sigma-a⁴⁹ weighting (T.J.R., unpublished) resulted in a model containing the entire DNA and histone octamer minus some of the tail regions (H3: amino acids 38–135; H3': 20–135; H4: 20–102; H4': 16–102; H2A: 4–118; H2A': 12–118; H2B and H2B': 24–122). After group temperature-factor refinement and application of a solvent mask, an *R*-factor of 22.4 (*R*_{free} of 30.2) was obtained. All amino acids included in the model were within the allowed regions of the Ramachandran plot. Tail regions generally had high temperature factors (>60), unless engaged in crystal contacts. Graphic figures were prepared with MidasPlus⁴⁷ and GRASP⁴⁸. The electron density displayed is at 1.3× sigma of the map.

Received 8 May; accepted 9 July 1997.

- Kornberg, R. D. Structure of chromatin. *Annu. Rev. Biochem.* **46**, 931–954 (1977).
- McGhee, J. D. & Felsenfeld, G. Nucleosome structure. *Annu. Rev. Biochem.* **49**, 1115–1156 (1980).
- Widom, J. Toward a unified model of chromatin folding. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 365–395 (1989).
- van Holde, K. E. *Chromatin* (Springer, New York, 1988).
- Blank, T. A. & Becker, P. B. The effect of nucleosome phasing sequences and DNA topology on nucleosome spacing. *J. Mol. Biol.* **260**, 1–8 (1996).
- Wallrath, L. L., Lu, Q., Granok, H. & Elgin, S. C. R. Architectural variations of inducible eukaryotic promoters: Preset and remodeling chromatin structures. *BioEssays* **16**, 165–170 (1994).
- Flaus, A., Luger, K., Tan, S. & Richmond, T. J. Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals. *Proc. Natl Acad. Sci. USA* **93**, 1370–1375 (1996).
- Travers, A. A. DNA bending and nucleosome positioning. *Trends Biochem. Sci.* **12**, 108–112 (1987).
- Flaus, A. & Richmond, T. J. Positioning and stability of nucleosomes on MMTV 3' LTR sequences. *J. Mol. Biol.* (in the press).
- Wasyluk, B. & Chambon, P. Transcription by eukaryotic RNA polymerases A and B of chromatin assembled *in vitro*. *Eur. J. Biochem.* **98**, 317–327 (1979).
- Grunstein, M. Histone function in transcription. *Annu. Rev. Cell Biol.* **6**, 643–678 (1990).
- Paranjape, S. M., Kamakaka, R. T. & Kadonaga, J. T. Role of chromatin structure in the regulation of transcription by RNA polymerase II. *Annu. Rev. Biochem.* **63**, 265–297 (1994).
- Polach, K. J. & Widom, J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J. Mol. Biol.* **258**, 800–812 (1996).
- Felsenfeld, G. Chromatin unfolds. *Cell* **86**, 13–19 (1996).
- Schild, C., Claret, F. X., Wahli, W. & Wolffe, A. P. A nucleosome-dependent static loop potentiates estrogen-regulated transcription from the *Xenopus* vitellogenin-B1 promoter *in vitro*. *EMBO J.* **12**, 423–433 (1993).
- Truss, M. *et al.* Hormone induces binding of receptors and transcription factors to a rearranged nucleosome on the MMTV promoter *in vivo*. *EMBO J.* **14**, 1737–1751 (1995).

- Rhodes, D., Brown, R. S. & Klug, A. *Meth. Enzymol.* **420–428** (Academic, San Diego, 1989).
- Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 Å resolution. *Nature* **311**, 532–537 (1984).
- Finch, J. T. *et al.* X-ray and electron microscope studies on the nucleosome structure. *FEBS Lett.* **51**, 193–197 (1979).
- Arents, G., Burlingame, R. W., Wang, B.-C., Love, W. E. & Moudrianakis, E. N. The nucleosomal core histone octamer at 3.1 Å resolution: A tripartite protein assembly and a left-handed superhelix. *Proc. Natl Acad. Sci. USA* **88**, 10148–10152 (1991).
- Richmond, T. J., Rechsteiner, T. & Luger, K. Studies of nucleosome structure. *Cold Spring Harbor Symp. Quant. Biol.* **LVIII**, 265–272 (1993).
- Richmond, T. J., Searles, M. A. & Simpson, R. T. Crystals of a nucleosome core particle containing defined sequence DNA. *J. Mol. Biol.* **199**, 161–170 (1988).
- Luger, K., Rechsteiner, T. J., Flaus, A., Waye, M. M. Y. & Richmond, T. J. Characterization of nucleosome core particles containing histone proteins made in bacteria. *J. Mol. Biol.* (in the press).
- Camerini-Otero, R. D. & Felsenfeld, G. Sulfhydryl modification of nucleosome. *Proc. Natl Acad. Sci. USA* **74**, 5519–5523 (1977).
- Harp, J. M. *et al.* X-ray diffraction analysis of crystals containing twofold symmetric nucleosome core particles. *Acta Crystallogr. D* **52**, 283–288 (1996).
- Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).
- Rhodes, D. & Klug, A. Sequence-dependent helical periodicity of DNA. *Nature* **292**, 378–380 (1981).
- Dickerson, R. E., Goodsell, D. S. & Neidle, S. "... The tyranny of the lattice..." *Proc. Natl Acad. Sci. USA* **91**, 3579–3583 (1994).
- Travers, A. A. & Klug, A. in *DNA Topology and its Biological Effects* (eds Cozzarelli, N. R. & Wang, J. C.) 57–106 (Cold Spring Harbor Press, Cold Spring Harbor, New York, 1990).
- Pryciak, P. M. & Varmus, H. E. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69**, 769–780 (1992).
- Pruss, D., Bushmann, F. D. & Wolffe, A. P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl Acad. Sci. USA* **91**, 5913–5917 (1994).
- Polach, K. J. & Widom, J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J. Mol. Biol.* **254**, 130–149 (1995).
- Studitsky, V. M., Clark, D. J. & Felsenfeld, G. Overcoming a nucleosomal barrier to transcription. *Cell* **83**, 19–27 (1995).
- Hirschhorn, J. N., Brown, S. A., Clark, C. D. & Winston, F. Evidence that SNF2/SWI2 and SNF5 activate transcription in yeast by altering chromatin structure. *Genes Dev.* **6**, 2288–2298 (1992).
- Kruger, W. *et al.* Amino acid substitutions in the structured domains of histones H3 and H4 partially relieve the requirement of the yeast SWI/SNF complex for transcription. *Genes Dev.* **9**, 2770–2779 (1995).
- Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc. Natl Acad. Sci. USA* **73**, 1897–1901 (1976).
- Thoma, F., Koller, T. & Klug, A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J. Cell Biol.* **83**, 403–427 (1979).
- Widom, J. & Klug, A. Structure of the 300 Å chromatin filament: X-ray diffraction from oriented samples. *Cell* **43**, 207–213 (1985).
- Graziano, V., Gerchman, S. E., Schneider, D. K. & Ramakrishnan, V. Histone H1 is located in the interior of the chromatin 30-nm filament. *Nature* **368**, 351–354 (1994).
- Hecht, A., Laroche, T., Strahl-Bolsinger, S., Gasser, S. M. & Grunstein, M. Histone H3 and H4 N-termini interact with SIR3 and SIR4 proteins: A molecular model for the formation of heterochromatin in yeast. *Cell* **80**, 583–592 (1995).
- Starich, M. R., Sandman, K., Reeve, J. N. & Summers, M. F. NMR structure of HmfB from the hyperthermophile, *Methanothermobacter fervidus*, confirms that this archaeal protein is a histone. *J. Mol. Biol.* **255**, 187–203 (1996).
- Xie, X. *et al.* Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature* **380**, 316–322 (1996).
- Yang, T. P., Hansen, S. K., Oishi, K. K., Ryder, O. A. & Hamkalo, B. A. Characterization of a cloned repetitive DNA sequence concentrated on the human X chromosome. *Proc. Natl Acad. Sci. USA* **79**, 6593–6597 (1982).
- O'Halloran, T. V., Lippard, S. J., Richmond, T. J. & Klug, A. Multiple heavy-atom reagents for macromolecular X-ray structure determination. Application to the nucleosome core particle. *J. Mol. Biol.* **194**, 705–712 (1987).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brünger, A. *X-PLOR v3.1 Manual* (Yale Univ. Press, New Haven, 1992).
- Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. The MIDAS display system. *J. Mol. Graph.* **6**, 13–27 (1988).
- Nicholls, A., Sharp, K. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
- Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42**, 140–149 (1986).
- Böhm, L. & Crane-Robinson, C. Proteases as structural probes for chromatin: the domain structure of histones. *Biosci. Rep.* **4**, 365–386 (1984).
- Goldknopf, I. L. & Busch, H. Isopeptide linkage between nonhistone and histone 2A polypeptides of chromosomal conjugated-protein A24. *Proc. Natl Acad. Sci. USA* **74**, 864–868 (1977).

Acknowledgements. We thank S. Halford for the EcoRV enzyme; A. Flaus, T. Rechsteiner and S. Tan for support and discussion; and C. Riekel and his co-workers at ID13 of the ESRF, Grenoble for support and cooperation. This research was supported in part by the Swiss National Fond.

Correspondence and requests for materials should be addressed to T.J.R. The atomic coordinates have been deposited in the Brookhaven Protein Databank under code laoi.

The radio afterglow from the γ -ray burst of 8 May 1997

D. A. Frail*, S. R. Kulkarni†, L. Nicastro‡, M. Feroci§ & G. B. Taylor*

* National Radio Astronomy Observatory, Socorro, New Mexico 87801, USA

† Division of Physics, Mathematics and Astronomy 105-24, Caltech, Pasadena, California 91125, USA

‡ Istituto Tecnologie e Studio Radiazioni Extraterrestri, CNR, I-40129 Bologna, Italy

§ Istituto di Astrofisica Spaziale, CNR, Via E. Fermi, 21, 00044 Frascati, Italy

Important insight into the nature of γ -ray bursts (GRBs) has been gained in recent months mainly due to the immediate, precise localization of the bursts^{1–3} and the discovery of relatively long-lived X-ray afterglows⁴ by the satellite BeppoSAX⁵. These advances have enabled deep searches which have led to the discovery of optical transients^{6,7} coincident with fading X-ray sources. Optical spectroscopy of the latest burst (GRB970508; ref. 8) has clearly demonstrated that it lies at a cosmological distance, thus resolving a long-standing controversy about the distance scale to GRBs. Here we report a variable radio source within the error box of GRB970508 and coincident with the optical transient. We suggest that this is the much-sought-after radio counterpart of a GRB. If the observed fluctuations in the radio emission ('twinkling') are a result of a strong scattering by the irregularities in the ionized Galactic interstellar gas, then the source must have an angular size of about 3 microarcseconds in the first few weeks. The damping of the fluctuations with time indicates that the source expands to a significantly larger size later on.

Radio observations began at the Very Large Array (VLA) at 1997 May 9.06 UT, only 3.7 h after the γ -ray burst². This and the next observation (May 9.84 UT) were conducted in the 1.43-GHz band. No radio sources were found in the preliminary 5-arcmin error circle, with 2σ limits of 0.09 and 0.18 mJy, respectively. On 1997 May 10 UT the position of the GRB was refined³ to a circle of radius 3 arcmin. At about the same time, an optical variable was reported⁷. This source lies within the 50-arcsec (radius) localization circle of a transient X-ray source, 1SAX J0653.8 + 7916 (ref. 4). The refined position of the burst enabled us to start observing at higher radio frequencies, specifically 4.86 and 8.46 GHz, for which the fields of view are correspondingly smaller. The first 8.46-GHz observations (1997 May 13.96 UT) resulted in the detection of a single source (flux density $S = 0.43 \pm 0.03$ mJy) within the 3-arcmin GRB error circle and coincident with the optical variable^{7,9}; we refer to this source as VLA J065349.4 + 791619.

Our monitoring programme showed that VLA J065349.4 + 791619 is quite variable; the observations are summarized in Table 1 and Fig. 1. The source is unresolved at our highest angular resolution of 0.8 arcsec; the fraction of linear polarization is $< 9\%$. The probability of detecting a background radio source at this level in the 50-arcsec-radius error circle by chance is 0.02, small but not insignificant¹⁰. However, the high degree of radio variability and the unusual radio spectrum (see below and Table 1) make this object interesting. Regardless of these considerations we note that the radio source coincides, within the experimental errors, with the optical variable⁹ which is now widely regarded as the optical afterglow of GRB970508. Thus we conclude that VLA J065349.4 + 791619 is most likely to be a radio source associated with GRB970508. Having accepted this hypothesis we now proceed with the interpretation of the radio observations.

The temporal behaviour of the radio flux density in the several months following the initial burst has not mirrored the simple

smooth power-law decay of the optical^{9,11}. On the contrary, during the first three weeks the radio light curve at 8.46 GHz (Fig. 1a) is erratic with large (50%) fluctuations. The amplitude of these modulations have since decreased, with $\sim 15\%$ variations seen out to at least post-burst day 55. The same trends are seen at 4.86 GHz (Fig. 1b), although the sampling is not as complete.

Leaving aside these large fluctuations we note that the mean radio flux rises to a constant value, the 'plateau value' of $\sim 600 \mu\text{Jy}$ (similar at both 4.86 and 8.46 GHz), remains steady for about 2 months after which there is a slow fall-off. From Fig. 1a we see that the time to rise to the plateau value is about 3 weeks after the burst at 4.86 GHz but only about 10 d at 8.46 GHz. This trend, a rise time which increases with decreasing frequency, continues to flux measurements at 1.43 GHz (Table 1). At this frequency, during the first months the source is either undetectable or weak, with flux $< 100 \mu\text{Jy}$. Subsequently the source slowly rises to a value of 500 μJy towards the end of our monitoring period. This is about equal to the plateau value that we noted at the two higher frequencies.

Given the limited number of frequencies (ν) at which the fluxes are measured we use a simple power-law model for the radio spectrum: flux $S_\nu \propto \nu^\alpha$ where α is the spectral index. We specifically discuss the spectrum around 16–17 May 1997 as a number of different frequency measurements are available at this epoch. From Table 1, we note that the spectral index between 1.43 and 8.46 GHz is ~ 1 . This is confirmed by a detection¹² at 15 GHz ($S_{15} = 1.57 \pm 0.25$ mJy). A turnover in the spectrum is required at higher frequencies in order to explain the apparent non-detections at 86 GHz with the BIMA and OVRO instruments^{13,14}.

Table 1 Radio observations of GRB970508

Epoch (UT)	Δt (d)	$S(1.43)$ (μJy)	$S(4.86)$ (μJy)	$S(8.46)$ (μJy)	α
1997 May 09.06	0.16	$< 90 \pm 45$			
1997 May 09.84	0.93	$< 180 \pm 90$			
1997 May 13.96	5.06			430 ± 25	
1997 May 15.13	6.23	100 ± 32	330 ± 33	610 ± 33	$+1.11 \pm 0.11$
1997 May 16.49	7.59			520 ± 18	
1997 May 18.85	9.95	$< 130 \pm 63$	480 ± 64	610 ± 51	$+0.43 \pm 0.16$
1997 May 22.10	13.20		410 ± 47	570 ± 42	$+0.59 \pm 0.14$
1997 May 22.49	13.59		360 ± 31	880 ± 33	$+1.61 \pm 0.09$
1997 May 24.48	15.58			550 ± 44	
1997 May 24.96	16.06			470 ± 43	
1997 May 28.85	16.95			480 ± 45	
1997 May 27.67	18.77		770 ± 54	500 ± 49	-0.78 ± 0.12
1997 May 29.10	20.20		1350 ± 55	1200 ± 47	-0.21 ± 0.06
1997 May 30.95	22.05		740 ± 46	810 ± 48	$+0.16 \pm 0.09$
1997 May 31.99	23.09	105 ± 50	550 ± 43	290 ± 46	-1.15 ± 0.18
1997 Jun. 02.29	24.39	135 ± 45	650 ± 44	720 ± 62	$+0.18 \pm 0.11$
1997 Jun. 02.41	24.51			940 ± 27	
1997 Jun. 02.90	25.00		540 ± 46	760 ± 45	$+0.62 \pm 0.10$
1997 Jun. 03.94	26.04		530 ± 51	670 ± 43	$+0.42 \pm 0.12$
1997 Jun. 05.97	28.07		850 ± 36	690 ± 28	-0.38 ± 0.06
1997 Jun. 09.74	31.84		470 ± 38	530 ± 38	$+0.22 \pm 0.11$
1997 Jun. 12.99	35.09	130 ± 22	640 ± 43	600 ± 39	-0.12 ± 0.09
1997 Jun. 14.73	36.83		590 ± 69	460 ± 66	-0.45 ± 0.19
1997 Jun. 17.92	40.02		400 ± 49	560 ± 40	$+0.61 \pm 0.14$
1997 Jun. 18.96	41.06		640 ± 55	630 ± 63	-0.03 ± 0.13
1997 Jun. 20.86	42.96		550 ± 36	800 ± 32	$+0.68 \pm 0.08$
1997 Jun. 22.93	45.03		425 ± 43	645 ± 38	$+0.75 \pm 0.12$
1997 Jun. 26.63	48.73		610 ± 70	680 ± 46	$+0.20 \pm 0.13$
1997 Jun. 28.68	50.78		500 ± 34	565 ± 36	$+0.22 \pm 0.09$
1997 Jul. 02.36	54.46		805 ± 38	805 ± 43	$+0.00 \pm 0.07$
1997 Jul. 04.59	56.69		605 ± 50	710 ± 49	$+0.29 \pm 0.11$
1997 Jul. 11.10	63.20		630 ± 32	725 ± 40	$+0.25 \pm 0.07$
1997 Jul. 14.61	66.71		585 ± 48	680 ± 46	$+0.27 \pm 0.11$
1997 Jul. 15.33	67.43		550 ± 31	630 ± 37	$+0.24 \pm 0.08$
1997 Jul. 18.44	70.54	305 ± 48	505 ± 62	575 ± 61	$+0.23 \pm 0.16$
1997 Jul. 18.64	70.74		535 ± 43	515 ± 41	-0.07 ± 0.11
1997 Jul. 22.55	74.65		385 ± 51	440 ± 51	$+0.24 \pm 0.18$
1997 Jul. 28.95	81.05	485 ± 45			
1997 Jul. 30.46	82.56		360 ± 70	455 ± 85	$+0.42 \pm 0.27$
1997 Aug. 04.01	87.11		370 ± 38	380 ± 42	$+0.05 \pm 0.15$

Column headings have meanings as follows. 'Epoch' is the start of each observation. ' Δt ' is the time elapsed since start of the burst. Columns headed by ' $S(1.43)$ ', ' $S(4.86)$ ' and ' $S(8.46)$ ' contain the radio flux densities and their 1σ errors for VLA J065349.4 + 791619 at 1.43 GHz, 4.86 GHz and 8.46 GHz, respectively. ' α ' is the spectral index (defined by $S_\nu \propto \nu^\alpha$) between 4.86 and 8.46 GHz.

Table 1 shows the temporal variation of α (measured between 4.86 and 8.46 GHz); this quantity shows wild variations during the first few weeks, the period during which the largest flux variations are seen. At later times, when the fluctuations in flux density have dampened, α has settled to a positive value near 0.2. This predicts that the plateau value of the flux at 1.43 GHz will be smaller than that at 8.46 GHz by a factor of 1.4.

We now interpret the radio data in the context of theoretical models proposed for GRBs. Cosmological models of GRBs—so-called ‘fireball’ models—consider the evolution of 10^{52} erg of energy (the energy implied¹⁵ by the minimum distance⁸ to GRB970508) that is deposited either impulsively or over timescales typical of the durations of γ -ray bursts. This energy is rapidly converted to kinetic energy which results in a spherical blast wave moving at relativistic speeds. Lorentz factors for the blast wave (Γ) between 300 and 1,000 are invoked. X-ray, optical and radio emission on a timescale longer than the durations of the GRBs (following the γ -ray burst) are generic predictions^{16–18} of the fireball models. The afterglow emission arises from synchrotron emission from the particles shocked by the blast wave. The X-ray and optical afterglow of GRB970228, especially its decay, appear to fit a simple formulation of this model¹⁹.

But contrary to the expectation of the simple fireball model, the optical and radio afterglow of GRB970508 shows a slow rise to a plateau value before a decay sets in. Furthermore, the radio plateau value is $\sim 600 \mu\text{Jy}$, which is an order of magnitude larger than the peak⁹ optical value of $50 \mu\text{Jy}$. Vietri²⁰ explains these observations by invoking a radial dependence of the density of ambient gas (circumstellar matter) into which the blast wave is expanding. Waxman¹⁵ attributes these to the presence of a low-energy tail of electrons in the post-shock gas (which certainly exist but are ignored in the simplest fireball model). We are in the process of carrying out a comparison of the data presented here using these two models; this analysis will be presented elsewhere.

The observed rapid radio flux variations (Fig. 1a; especially in the first three weeks) and the accompanying dispersion in the spectral-index behaviour find no explanation in fireball models. This is

simple to understand as temporal evolution of the fireball is determined by the bulk Lorentz factor, Γ , which smoothly decreases with time. An independent observation providing evidence against intrinsic variations is that there are no reported measurements of such deep and rapid variations at optical wavelengths. Thus to understand these observed variations one must appeal to extrinsic causes.

Goodman²¹ has called attention to the potential role of interstellar scattering (ISS) in the variation of the observed radio fluxes. Irregularities in the interstellar medium scatter the incoming radio signal which then results in ‘twinkling’ of radio sources. There are two basic regimes of ISS²²: weak and strong. The relevant physical parameters are the Fresnel scale, r_F (the radius of the first Fresnel zone at the interstellar scattering layer, typically assumed to be located at distance $d \approx 1$ kpc from the observer) and the spatial coherence scale of the irregularities, r_c . When $r_c \gg r_F$, the effects of scattering are weak and therefore modest fluctuations in the flux are expected to be observed. In the strong scattering regime ($r_c \ll r_F$) scattering takes place owing to inhomogeneities on scales of r_c (diffractive) and also on much larger scales, r_F^2/r_c (refractive). However, refractive scattering results in modest flux variations and is broad-band in nature.

The observed deep modulation that is seen in the 8.46-GHz data (Fig. 1a), and the wild swings in spectral index, bear the signature of diffractive scintillation. If we accept this conclusion then θ_s , the angular size of the source, must be less than the ‘diffractive’ angle, $\theta_d = r_c/d$. For typical ISS parameters²¹ and an observing frequency of 8.46 GHz, $\theta_d \approx 3$ microarcseconds (μas). From observations of pulsars it is known that the irregularities in the Galactic interstellar medium vary from one line-of-sight to another. For this reason, we point out that the precision of the value quoted for θ_d is unlikely to be better than a factor of two. The minimum redshift of GRB970508 is $z = 0.835$ (ref. 8). Taking a value of $60 \text{ km s}^{-1} \text{ Mpc}^{-1}$ for the Hubble constant, we find a distance $D \approx 10^{28} \text{ cm}$. Thus the linear size of the source must be less than $R \approx 10^{17} \text{ cm}$.

The variations in the flux of the source appear to moderate after the first month. In the framework of ISS, the source has expanded to

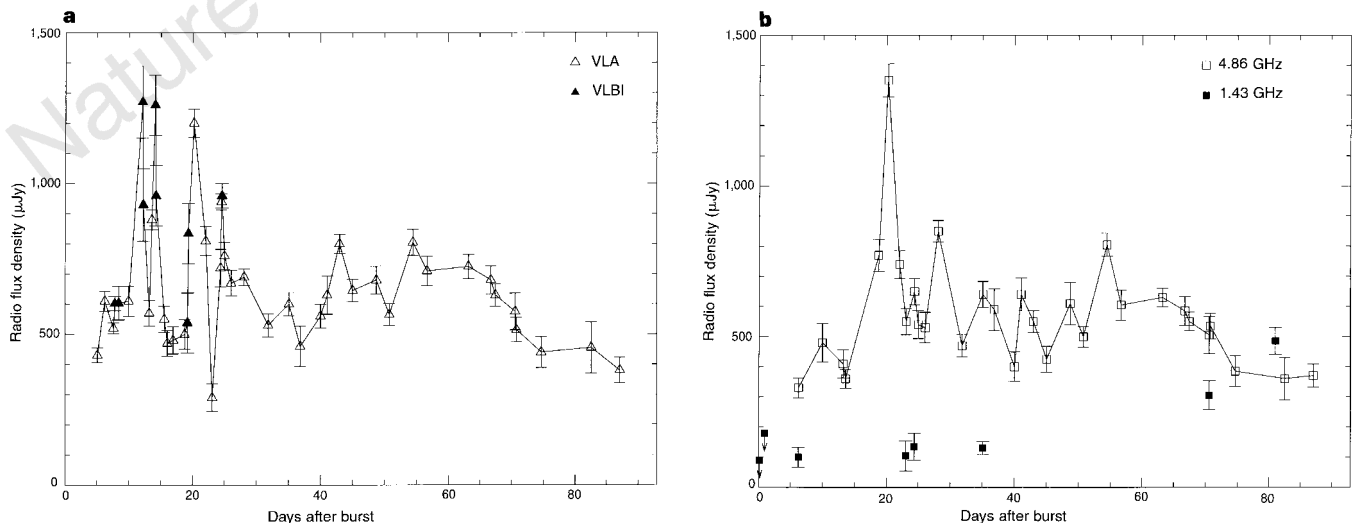


Figure 1 Light curves of the radio counterpart of GRB970508. **a**, Light curve at 8.46 GHz. Data points consist of VLA (open triangles) and VLB data (filled triangles)²³. The flux densities from the VLB measurements²⁵ were determined by splitting each data set in half and measuring them independently. There is close agreement between those flux densities which were independently measured by the VLA and VLB. The first detection of the source was made on May 13.96 UT, 5 days after the GRB event. The delay, if any, between the onset of the radio emission and the high-energy radiation is not well constrained, as the earliest VLA

observations were made at 1.43 GHz. Unfortunately, as shown in **b** and Table 1, the source was weak for the first month at this frequency. **b**, Light curve at 4.86 GHz (open squares) and 1.43 GHz (filled squares). The sampling at 4.86 GHz is sparser than at 8.46 GHz but the general characteristics of the light curve remain the same, with a plateau or a slight rise in the mean flux density and occasional ‘flares’. At 1.43 GHz the source remained weak or undetectable for the first month but has since risen to a level comparable to the plateau values seen at 4.86 and 8.46 GHz.

a large enough size that diffractive scattering is quenched. Goodman²¹ points out that refractive scintillation will dominate the scattering. For the parameters discussed above (8 GHz, typical high-latitude line-of-sight) the Fresnel angle, r_F/d , is coincidentally close to θ_d . When the fluctuations are dying down, the source must have a size comparable to θ_d . Goodman²¹ shows that in this regime the modulation index (the r.m.s. of the fluctuations to the mean value) is $\sim 12\%$ —precisely what is observed. Moreover, the fluctuations are expected to be broad-band as is indeed the case (see Fig. 1a). Consistent with this discussion, very long baseline (VLB) observations²³ show that the size of the source is $< 260 \mu\text{as}$.

We recognize that the above conclusions rest crucially on the assumption that the flux variations are induced by ISS. We now discuss additional evidence for (or against) this assumption. In the ISS framework, flux will vary on a timescale determined by the crossing time of the irregularities, $t_{\text{ISS}} \approx r_c/v_E = d\theta_d/v_E$ where v_E is the velocity of the irregularities with respect to the Earth. The differential delay between rays converging to the observer is $d\theta_d^2/c$ and thus the bandwidth of a constructive (or destructive) interference is $\Delta\nu_{\text{ISS}} \approx c/2\pi d\theta_d^2$. For our nominal value, $d = 1 \text{ kpc}$, $\theta_d = 3 \mu\text{as}$ (in the 8-GHz band) and $v_E \approx 30 \text{ km s}^{-1}$, we obtain $t_{\text{ISS}} = 4 \text{ h}$ and $\Delta\nu = 7 \text{ GHz}$. The VLA data presented here (Table 1) are too poorly sampled to measure t_{ISS} . However, the VLBA data presented by Taylor *et al.*²³ are of longer duration. A 25% variation is seen over 5 h which suggests that $t_{\text{ISS}} < 1 \text{ d}$. The wild variations seen in α (between 4.86 and 8.46 GHz; Table 1) requires that $\Delta\nu \approx 2 \text{ GHz}$. These are not severe disagreements. Both the definitions and the estimates for the two parameters discussed above are approximate. Further improvement in a quantitative sense is best done by carrying out ISS observations of pulsars and compact objects in this part of the sky and obtaining estimates of θ_d , d and v_E in a consistent manner.

An independent estimate for the size comes from an argument presented by Katz and Piran²⁴. We have had noted above that in the first month VLA J065349.4 + 791619 is essentially undetectable at 1.4 GHz but strong at higher frequencies. A simple interpretation is that VLA J065349.4 + 791619 is optically thick at 1.4 GHz in the first month. If so, the spectrum expected below the self-absorption frequency is²⁴;

$$S_\nu = 2\pi\nu^2 m_p \zeta (1+z) R^2 / D^2$$

where ζ is a parameter which describes the degree of electron equipartition in the shocked gas, and m_p is proton mass. Fireball models usually (for example, ref. 15) adopt a value of 0.1 for this parameter. Using the measured value of the flux at 1.43 GHz (Table 1), Katz and Piran²⁴ obtain $R \approx 10^{17} \text{ cm}$.

The interpretation of R defined above is unfortunately not a simple matter. What is more interesting is that the above equation predicts that the flux at 1.4 GHz should increase as θ_s^2 . Earlier we argued that the source must have significantly expanded between the first and the last month of observations. In accordance with our expectation, we note that the 1.43-GHz flux has increased from $< 100 \mu\text{Jy}$ to $500 \mu\text{Jy}$ during this period. Taken at face value, this suggests that θ_s has doubled in size over this period. This conclusion is in accordance with our assertion (from interstellar scintillation arguments) that the source has expanded considerably.

We have reported above the first measurement of the linear size of a fireball. From this we can extract the mean Lorentz factor of the expansion. Taking the above-mentioned size of $3 \mu\text{as}$ for the fireball at about $t = 2$ weeks post-burst, we can obtain a mean speed of the burst, R/t , of $4c$. For a relativistically expanding fireball, $R = f\gamma^2 ct$ and the apparent size is R/γ . Here γ is the Lorentz factor for the expanding shell, and the factor f depends on the dynamical details of the expanding fireball and varies between 2 and 14 (ref. 24). Thus $f\gamma \approx 4$, that is, the shell is mildly relativistic two weeks following the burst. $\square$

Received 8 August; accepted 28 August 1997.

- Costa, E. *et al.* Discovery of an X-ray afterglow associated with the γ -ray burst of 28 February 1997. *Nature* **387**, 783–785 (1997).
- Costa, E. *et al.* *IAU Circ.* No. 6649 (1997).
- Heise, J. *et al.* *IAU Circ.* No. 6654 (1997).
- Piro, L. *et al.* *IAU Circ.* No. 6656 (1997).
- Boella, G. *et al.* BeppoSAX, the wideband mission for X-ray astronomy. *Astron. Astrophys. Suppl. Ser.* **122**, 299–307 (1997).
- van Paradijs, J. *et al.* Transient optical emission from the error box of the γ -ray burst of 28 February 1997. *Nature* **386**, 686–689 (1997).
- Bond, H. E. *et al.* *IAU Circ.* No. 6665 (1997).
- Metzger, M. R. *et al.* Spectral constraints on the redshift of the optical counterpart to the γ -ray burst of 8 May 1997. *Nature* **387**, 878–880 (1997).
- Djorgovski, S. G. *et al.* The optical counterpart to the γ -ray burst GRB970508. *Nature* **387**, 876–878 (1997).
- Windhorst, R. A. *et al.* Microjansky source counts and spectral indices at 8.44 GHz. *Astrophys. J.* **405**, 498–517 (1993).
- Pian, E. *et al.* Hubble space telescope imaging of the optical counterpart to the gamma-ray burster GRB970508. *Astrophys. J.* (submitted).
- Pooley, G. & Green, D. A. *IAU Circ.* No. 6670 (1997).
- Smith, I. A. *et al.* *IAU Circ.* No. 6663 (1997).
- Shepherd, D. S., Metzger, M. R. & Kulkarni, S. R. *IAU Circ.* No. 6664 (1997).
- Waxman, E. γ -ray burst after-glow: confirming the cosmological fireball model. *Nature* (submitted).
- Mészáros, P. & Rees, M. J. Optical and long wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476**, 232–237 (1997).
- Paczynski, B. & Rhoads, J. E. Radio transients from gamma-ray bursters. *Astrophys. J.* **418**, L5–L8 (1993).
- Katz, J. I. Low frequency spectra from gamma-ray bursts. *Astrophys. J.* **432**, L107–L109 (1994).
- Wijers, R. A. M. J., Rees, M. J. & Meszaros, P. Shocked by GRB 970228: the afterglow of a cosmological fireball. *Mon. Not. R. Astron. Soc.* (submitted); Preprint astro-ph/9704158 (1997).
- Vietri, M. The afterglow of gamma ray bursts II: the cases of GRB 970228 and GRB 970508. *Astrophys. J.* (submitted).
- Goodman, J. Radio scintillation of gamma-ray burst afterglows. Preprint astro-ph/9706084 (1997).
- Narayan, R. in *Pulsars as Physics Laboratories* (eds Blandford, R. D., Hewish, A., Lyne, A. G. & Mestel, L.) 151–165 (Oxford Univ. Press, 1993).
- Taylor, G. B., Frail, D. A., Beasley, A. J. & Kulkarni, S. R. Position and parallax of the γ -ray burst of 8 May 1997. *Nature* **389**, 263–265 (1997).
- Katz, J. I. & Piran, T. Persistent counterparts of GRBs. Preprint astro-ph/9706141 (1997).

Acknowledgements. The VLA is a facility of the NSF operated under cooperative agreement by Associated Universities, Inc. D.A.F. and S.R.K. thank NRAO, and in particular B. Clark and M. Goss, for their support of the GRB programme. We thank R. Blandford and E. Waxman for discussions. S.R.K. is supported by the NSF and NASA.

Correspondence should be addressed to D.A.F. (e-mail: dfrail@nrao.edu).

Position and parallax of the γ -ray burst of 8 May 1997

G. B. Taylor*, D. A. Frail*, A. J. Beasley* & S. R. Kulkarni†

* National Radio Astronomy Observatory, Socorro, New Mexico 87801, USA

† Division of Physics, Mathematics and Astronomy 105-24, Caltech, Pasadena, California 91125, USA

γ -ray bursts (GRBs) have puzzled astronomers for almost three decades, primarily owing to the lack of identifications at other wavelengths. The detection of a radio counterpart¹ enables application of the powerful technique of very long baseline interferometry (VLBI) to this intriguing class of objects. Here we present VLBI monitoring of VLA J065349.4 + 791619 obtained between 8 and 25 days after the initial γ -ray burst. The radio emission is found to be very compact, with an angular extent of less than 1 milliarcsecond. We derive a position for the radio counterpart accurate to 200 microarcseconds, and constrain the proper motion of VLA J065349.4 + 791619 to be less than 50 milliarcseconds per year. We place an upper limit on the annual parallax of 1 milliarcsecond. These results are entirely consistent with the expectations of cosmological models.

On 16 May 1997 UT we started a programme of observations of the radio counterpart to GRB970508, VLA J065349.4 + 791619, with the 10-antenna Very Long Baseline Array (VLBA) spanning the United States. The first observation was conducted 7.8 d after the γ -ray burst and also included the 100-m reflector at Effelsburg, Germany and the Very Large Array (VLA). All the observations were conducted at a central frequency of 8.41 GHz. Signals in both senses of circular polarization were recorded on magnetic tapes and

Table 1 Gaussian model fits

Epoch	<i>t</i> (h)	<i>S</i> (μ Jy)	r.m.s. (μ Jy)	Δ RA (μ as)	Δ Dec. (μ as)	FWHM (μ as)	Restoring beam (μ as)	Antennas
May 16.7	10.5	602	24	-239	286	487	1,060, 840, 23°	BEFHKL MNOPSY
May 17.3	5.6	604	56	15	-63	408	1,920, 910, 56°	BFHLMNOPS
May 21.1	5.6	1,166	75	139	-63	709	1,880, 940, -75°	BFHKNOPS
May 23.2	5.6	1,173	63	179	49	1,083	2,310, 1,450, -78°	BFHKL NOPS
May 28.2	5.6	683	53	-184	-51	263	1,850, 940, -86°	BFHKL MNOPS
Jun. 2.5	5.6	957	40	87	-157	406	2,090, 1,100, -51°	BFHKL MNOPSY

Column headings have meanings as follows. 'Epoch', mean epoch of the observations. Year is 1997. '*t*', total length of observing run. '*S*', flux density at 8.4 GHz. Calibration was done with respect to noise tubes of known value and measured antenna gains. We find good agreement between the flux density measured by the VLBA and that derived from nearly contemporaneous (and in two cases, simultaneous) VLA measurements at the same frequency. 'r.m.s.', the 1σ noise in the flux per beam in the image. ' Δ ', offset in right ascension and declination from the mean position. The error in these offsets is dominated by unmodelled tropospheric fluctuations which result in a systematic error (determined by observations of a second calibrator) of 140μ as in each coordinate (1σ). 'FWHM', full-width at half-maximum of a gaussian component fit to the visibility data. 'Restoring beam' a two-dimensional gaussian with the following three parameters: the FWHM of the major axis, the FWHM of the minor axis, and the position angle of the major axis of the elliptical beam (the angle measured from north round to east). 'Antennas', those antennas successfully participating in the observing run. B, Brewster; E, Effelsberg 100-m; F, Fort Davis; H, Hancock; K, Kitt Peak; L, Los Alamos; M, Mauna Kea; N, North Liberty; O, Owens Valley; P, Pie Town; S, St Croix; Y, phased VLA.

later processed at the VLBA correlator in Socorro, New Mexico. Table 1 gives further details of the observations and a summary of the results.

In all six observations (Fig. 1), VLA J065349.4 + 791619 is unresolved, although the final resolution of the instrument depends on the antennas involved and the degree to which atmospheric phase fluctuations could be removed (the latter effect is somewhat analogous to the phenomenon of optical 'seeing'). Under the best observing conditions (May 28.4 UT) we place a limit on the source size of $\Delta\theta < 0.26$ milliarcsecond (mas). The corresponding lower limit on the brightness temperature $T_B = S\lambda^2/2k\Delta\Omega$ is 1.4×10^8 K; here $\lambda = 3.6$ cm is the wavelength of observation, k is the Boltzmann constant and $\Delta\Omega = \pi\Delta\theta^2/4$ is the solid angle of the source. The deepest image (May 16.7 UT) allows us to limit the extended emission on scales 10 mas–20 arcsec to less than $\sim 50\mu$ Jy. The mean position for VLA J065349.4 + 791619 (obtained by averaging the positions given in Table 1) is right ascension (RA) 06 h 53 min 49.45131 s, declination (dec.) $+79^\circ 16' 19.51312''$. The 1σ error in RA is 170 microarcseconds (μ as) and 150 μ as in dec. The

mean position is epoch 1997.45 and equinox J2000. This position is not an absolute position but referenced to the calibrator, 0718 + 793, whose position is known to an accuracy of 20 μ as (ref. 2).

The radio emission may originate with the explosive event that produced the γ -ray, X-ray and optical transients, or it may be associated with the host galaxy of GRB970508. Here we compare the radio properties of VLA J065349.4 + 791619 to those of both 'normal' and 'active' galaxies, before going on to explore the predictions of cosmological fireball models.

At the sub-millijansky level, extragalactic sources are identified with host galaxies having a median optical magnitude $V = 23$ (ref. 3). The median angular size in the radio of these sources is 2.6 arcsec, and in 90% of the population the radio emission is thought to be a combination of steep-spectrum synchrotron emission from a galactic disk and thermal-bremsstrahlung from large-scale star formation³. About 10% of the sub-millijansky population may be dominated by emission from a nuclear starburst or a weak active galactic nucleus (AGN). Our VLBI brightness temperature

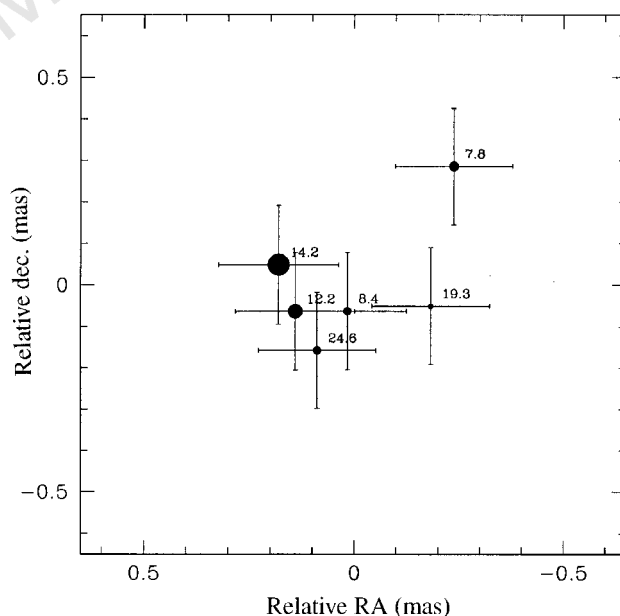


Figure 1 Positions of VLA J065349.4 + 791619 from the first six VLBI epochs. (RA, right ascension; dec., declination) The zero point was selected to be the average position reported in the text. The symbol size is proportional to the upper limit on the source size derived from each observation (Table 1), and the time elapsed in days since GRB970508 was detected by BeppoSAX is shown next to each measurement. The atmosphere affects the phase of the signal from the source differently at each of these widely spaced antennas. To overcome this and other instrumental contributions we alternated observations of VLA J065349.4 + 791619 with an observation of the strong, compact 'calibrator' source B0718 + 793

located 1.5° away. This technique, known as phase-referencing¹⁴, is now routinely used for observations of faint sources with the VLBA. The switching time for the first and sixth epochs (involving the phased VLA) was 2 min and 1.5 min for all other epochs. We fitted a linear model to the data and place an upper limit on the proper motion of 12 ± 15 mas yr⁻¹. The position measurements took place over a month and we see no systematic trend in the apparent position. Thus we place a rough upper limit of a milliarcsecond on the annual trigonometric parallax to the source. This limit requires the distance of the radio source to be > 1 kpc.

limit of $>10^8$ K rules out galactic disks and star formation as the source of the radio emission from VLA J065349.4 + 791619.

Next we consider the possibility that the radio emission from VLA J065349.4 + 791619 originates from the central 'engine' of an AGN. The radio power of the source at a rest frequency of 17 GHz (for an assumed redshift of 1) is 10^{24} W Hz⁻¹. This is comparable to the radio powers of nearby low-luminosity AGNs⁴. These radio sources, however, are generally dominated by emission from extended radio lobes. In the few cases where the cores are sufficiently strong, VLBI observations have revealed considerable structure⁵. In most cases this structure has a surface brightness a factor of 20–80 below that of the peak, so it is difficult to make a direct comparison to VLA J065349.4 + 791619 as (at best) the r.m.s. noise level is a factor of 20 below the peak brightness. The radio spectrum of VLA J065349.4 + 791619 (ref. 1), on average, resembles that of the so-called "GHz-peaked" radio sources. However, these sources show little variability and have complicated structure on the parsec scale^{6,7}. Gravitationally lensed sources at similar redshifts may have a comparable intrinsic radio power, but these also show much less radio variability than GRB970508, from 0 to 10%. The morphologies of the gravitationally lensed sources are predominantly core plus jet^{8,9}. The extreme variability of VLA J065349.4 + 791619 (ref. 1) is characteristic of a class of radio sources known as intraday variables, but these originate in flat-spectrum radio sources such as BL Lacs and quasars¹⁰.

Optical observations of the galaxy in which this radio source is located are also relevant. The optical source continues to fade and as yet there are no indications of any host galaxy to a limit of R-band magnitude $R \approx 24.5$ mag (ref. 11). This rules out typical radio galaxies as the host galaxy of these objects would be bright.

We conclude that GRB970508 is unlikely to be either radio emission from a normal galaxy or radio emission from the nucleus of an active galaxy. This conclusion is quite interesting as there are some models which posit that GRBs are beamed nuclear sources. The evidence presented above does not offer firm proof that this is not the case, but renders such a hypothesis less likely. Thus we now arrive at the preferred interpretation: GRB970508 is the radio counterpart or afterglow of the γ -ray burst of 8 May 1997 located in a faint galaxy.

Proceeding with the hypothesis that VLA J065349.4 + 791619 is the radio afterglow from GRB970508 we note that our observations rule out significant gravitational lensing on scales of 10 mas–300 arcsec. This covers the range of lenses discussed in the literature and rules out any substantial flux magnification of the γ -ray burst by gravitational lensing.

According to theoretical estimates of the angular diameter of the fireball¹², a future burst with the same radio luminosity as GRB970508 but at a redshift of 0.2 could be resolvable by VLBI observations. This would open the way for using VLBI to study the detailed evolution of a resolved GRB in much the same way as has recently been done for extragalactic supernovae¹³. □

Received 8 August; accepted 28 August 1997.

1. Frail, D. A., Kulkarni, S. R., Nicastro, L., Feroci, M. & Taylor, G. B. The radio afterglow from the γ -ray burst of 8 May 1997. *Nature* **389** 261–263 (1997).
2. Johnston, K. J. *et al.* A radio reference frame. *Astron. J.* **110**, 880–915 (1995).
3. Windhorst, R. A., Fomalont, E. B., Partridge, R. B. & Lowenthal, J. D. Microjansky source counts and spectral indices at 8.44 GHz. *Astrophys. J.* **405**, 498–517 (1993).
4. Giovannini, G., Feretti, L. & Comoretto, C. VLBI observations of a complete sample of radio galaxies. I. Snapshot data. *Astrophys. J.* **358**, 159–163 (1990).
5. Lara, L. *et al.* VLBI observations of a complete sample of radio galaxies. VII. Study of the FR I Sources 3C 31, 4C 35.03 and 3C 264. *Astrophys. J.* **474**, 179–187 (1997).
6. Taylor, G. B., Readhead, A. C. S. & Pearson, T. J. Pinpointing the center of activity in compact symmetric objects. I. Sources from the Pearson-Readhead survey. *Astrophys. J.* **463**, 95–104 (1996).
7. O'Dea, C. P., Stanghellini, C., Baum, S. A. & Charlot, S. On the host galaxies of the gigahertz peaked-spectrum radio sources. *Astrophys. J.* **470**, 806–813 (1996).
8. Campbell, R. M., Lehar, J., Corey, B. E., Shapiro, I. I. & Falco, E. E. VLBI observations of the gravitational lens system 0957 + 561. *Astron. J.* **110**, 2566–2569 (1995).
9. Porcas, R. W. in *Radio Emission from Galactic and Extragalactic Compact Sources* (eds Zensus, J. A., Wrobel, J. M. & Taylor, G. B.) (IAU Colloq. 164, Astron. Soc. Pacific, San Francisco, in the press).
10. Wagner, S. J. & Witzel, A. Intraday variability in quasars and BL Lacs objects. *Annu. Rev. Astron. Astrophys.* **33**, 163–197 (1995).

11. Pian, E. *et al.* Hubble space telescope imaging of the optical transient associated with GRB970508. *Astrophys. J.* (submitted).
12. Sari, R. Hydrodynamics of GRB afterglow. Preprint astro-ph/9706078 (1997).
13. Marcaide, M. *et al.* Expansion of SN 1993J. *Science* **270**, 1475–1478 (1995).
14. Beasley, A. J. & Conway, J. E. in *Very Long Baseline Interferometry and the VLBA* (eds Zensus, J. A., Diamond, P. J. & Napier, P. J.) 327–341 (ASP Conf. Ser. Vol. 82, Astron. Soc. Pacific, San Francisco, 1995).

Acknowledgements. The VLA and VLBA are facilities of the NSF operated under cooperative agreement by Associated Universities, Inc. We thank the VLBA, VLA and Effelsberg operational staff, and the VLBA correlator staff, for their extraordinary efforts on behalf of this project. S.R.K. is supported by the NSF and NASA.

Correspondence should be addressed to G.B.T. (e-mail: gtaylor@nrao.edu).

Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite

M. H. Engel* & S. A. Macko†

* School of Geology and Geophysics, The University of Oklahoma, Norman, Oklahoma 73019, USA

† Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia 22903, USA

Many amino acids contain an asymmetric centre, occurring as laevorotatory, L, or dextrorotatory, D, compounds. It is generally assumed that abiotic synthesis of amino acids on the early Earth resulted in racemic mixtures (L- and D-enantiomers in equal abundance). But the origin of life required, owing to conformational constraints, the almost exclusive selection of either L- or D-enantiomers^{1,2}, and the question of why living systems on the Earth consist of L-enantiomers rather than D-enantiomers is unresolved³. A substantial fraction of the organic compounds on the early Earth may have been derived from comet and meteorite impacts^{4–6}. It has been reported previously that amino acids in the Murchison meteorite exhibit an excess of L-enantiomers⁷, raising the possibility that a similar excess was present in the initial inventory of organic compounds on the Earth. The stable carbon isotope compositions of individual amino acids in Murchison support an extraterrestrial origin⁸—rather than a terrestrial overprint of biological amino acids—although reservations have persisted (see, for example, ref. 9). Here we show that individual amino-acid enantiomers from Murchison are enriched in ¹⁵N relative to their terrestrial counterparts, so confirming an extraterrestrial source for an L-enantiomer excess in the Solar System that may predate the origin of life on the Earth.

The Murchison meteorite is the most recently observed fall of an organic-rich, Type CM meteorite. Engel and Nagy⁷ reported an L-enantiomer excess for several amino acids in Murchison. The exotic amino-acid distribution, notably the absence or trace level of many common protein amino acids (for example, serine, threonine, methionine, tyrosine, phenylalanine, histidine, lysine, arginine), indicated that the L-excess for amino acids detected (for example, alanine, glutamic acid, aspartic acid) was not likely to be the result of contamination after impact. These results were incongruous with the initial report of approximately racemic amino acids in another Murchison stone¹⁰. Consequently, criteria independent of amino-acid distribution and stereochemistry have been sought to determine the source of the L-excess.

Stable isotopes have been used to assess the origins of organic matter in extraterrestrial materials (see, for example, refs 11–13). For instance, stable-isotope values ($\delta^{13}\text{C}$, $\delta^{15}\text{N}$, δD) of organic extracts of the Murchison meteorite are moderately-to-highly

enriched in the heavy stable isotopes ($^{13}\text{C} < ^{15}\text{N} < \text{D}$)^{11,14,15}. Similar isotope enrichments for interstellar clouds indicate that this is a likely source for the precursors of the Murchison amino acids (see, for example, ref. 9). In contrast, terrestrial organic matter of biogenic origin on Earth is usually moderately-to-substantially depleted in ^{13}C , ^{15}N and D relative to interstellar values, a consequence of isotope fractionations associated with biosynthetic processes such as photosynthesis¹⁶.

To date, most stable-isotope analyses of Murchison's organic matter have been determined on bulk extracts by conventional methods (see, for example, ref. 11). However, bulk organic extracts are comprised of mixtures of compounds, each having a potentially distinct isotopic composition. Thus, the stable-isotope values for bulk samples are weighted average values for the mixtures. In the few instances where $\delta^{13}\text{C}$ values have been determined for individual compounds (for example, carboxylic acids, amino acids, hydrocarbons)^{8,14,17,18}, differences in isotope compositions suggest distinct fractionations resulting from the apparent synthesis of higher-molecular-weight compounds from lower-molecular-weight homologues.

We previously attempted to determine whether the L-enantiomer excess in Murchison was extraterrestrial by using gas chromatography/combustion/isotope-ratio mass spectrometry (GC/C/IRMS) to obtain $\delta^{13}\text{C}$ values of individual amino-acid enantiomers. The results, although encouraging, were not definitive, because the $\delta^{13}\text{C}$ values for some components were only slightly-to-moderately enriched in ^{13}C relative to terrestrial organic matter. The $\delta^{15}\text{N}$ value for a bulk amino-acid extract of Murchison was substantially enriched (+90‰)¹¹ relative to biological material (−10 to +20‰). Thus, establishing the $\delta^{15}\text{N}$ values of individual amino acids should provide a better test of the hypothesis that the L-enantiomer excess in Murchison is extraterrestrial in origin.

A 7.3-g interior sample of a 103-g Murchison meteorite stone from the R. A. Langheinrich collection was ground to fine powder and extracted in distilled water for 8 h at 100 °C. This water extract was evaporated to dryness and then hydrolysed in 6M HCl for 24 h at 100 °C. The amino-acid fraction was isolated by cation-exchange chromatography as previously reported⁷. A portion of the amino-acid extract was analysed by high performance liquid chromatography (HPLC) to determine amino-acid abundances¹⁹. A second aliquot was derivatized to trifluoroacetyl isopropyl esters to determine stereochemistry and compound purity by gas chromatography (GC) and gas chromatography/mass spectrometry (GC/MS) as well as stable nitrogen isotope compositions of the individual stereoisomers by GC/C/IRMS. Specific details of the modifications to the GC/C/IRMS carbon-isotope methodology, as well as assessments of precision on standard materials are reported elsewhere^{20,21}. Briefly, derivatization of amino acids for stable nitrogen isotope analysis by GC/C/IRMS results in minimal isotopic fractionation and no additional nitrogen is added to the amino acid during the acylation and esterification steps²². Thus, unlike stable carbon isotope analysis of amino acids by GC/C/IRMS^{23,24}, no corrections are required to obtain the $\delta^{15}\text{N}$ values of the underivatized amino acids. Bulk $\delta^{15}\text{N}$, $\delta^{13}\text{C}$ analyses of acidified portions of the meteorite were accomplished using an off-line preparation procedure involving high-temperature combustion in quartz, cryogenic gas purification and isotope analysis on a PRISM IRMS. Carbonate isotope abundances ($\delta^{13}\text{C}$, $\delta^{18}\text{O}$) on a non-acidified portion of the stone were determined on a PRISM Micromass automated carbonate system.

The $\delta^{13}\text{C}$ and $\delta^{15}\text{N}$ values for the acidified bulk sample of this Murchison stone are −11.3‰ (with respect to the PDB standard) and +30.1‰ (with respect to atmospheric N_2), respectively, and are in agreement with previous findings⁸. Stable carbon and oxygen isotope values for the carbonate component of this stone are +38.6‰ (PDB) and +29.9‰ (SMOW standard), respectively, within the range of values reported for other Murchison stones²⁵.

Amino-acid distributions and abundances (Table 1) are also in good agreement with previous findings (for example, refs 8 and 26).

Owing to the need for adequate separation between individual amino acids eluting from the gas chromatographic column, not all amino acids or both stereoisomers were able to be isotopically assessed. For amino acids that could be analysed (Table 1), the $\delta^{15}\text{N}$ values are enriched in ^{15}N and fall well out of the average range of values reported for terrestrial organic matter (−10 to +20‰). With the notable exceptions of α -aminoisobutyric acid, sarcosine and glycine, the remaining amino acids have similar $\delta^{15}\text{N}$ values, implying a single, possibly interstellar¹⁵, homogeneous source for this nitrogen. The extreme ^{15}N enrichment observed for α -aminoisobutyric acid and sarcosine indicates either a unique precursor source or perhaps isotope fractionation during its formation or subsequent decomposition. The more depleted value for glycine may be associated with processes of formation or destruction of this or other amino acids, as glycine is known to form by the decom-

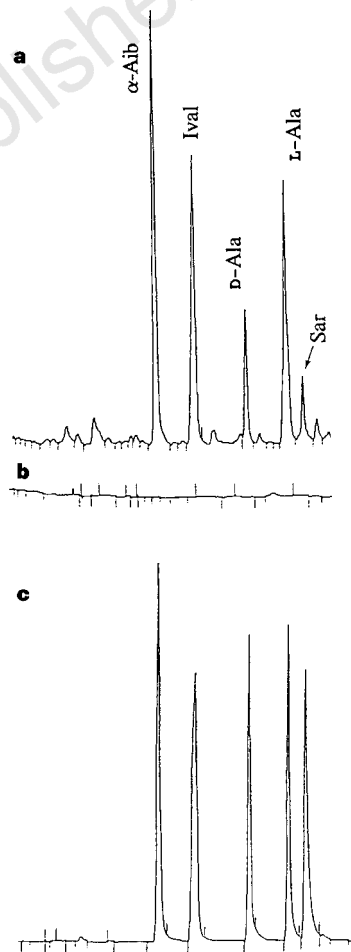


Figure 1 **a**, Gas chromatogram showing the resolution of D-alanine and L-alanine in the Murchison meteorite. **b**, Procedural blank chromatogram that preceded the meteorite analysis. **c**, Chromatogram of a standard mixture of the amino acids. Abbreviations are as follows: α -Aib, α -aminoisobutyric acid; Ival, isovaline; D-Ala, D-alanine; L-Ala, L-alanine; Sar, sarcosine. For GC, GC/MS and GC/C/IRMS, the separation of the trifluoroacetyl isopropyl esters of the amino acid stereoisomers was accomplished using a Chirasil-Val 50 m $\times$ 0.25 mm (internal diameter) fused-silica capillary column (Alltech, Deerfield, IL). The GC/C/IRMS consists of a Hewlett Packard 5890 GC coupled to an OPTIMA (Micromass, Manchester, UK) stable-isotope mass spectrometer through a cupric oxide/nichrome wire oxidation furnace, a copper reduction furnace and a liquid-nitrogen carbon dioxide/water cold trap²¹. The GC and GC/MS analyses were performed using a Hewlett Packard 5890A GC and a Hewlett Packard GC/MSD.

position of other amino acids²⁷. Explanations for the abundances beyond this first report will require additional experimentation and analysis; in particular, establishing the $\delta^{15}\text{N}$ values of additional components in this complex mixture.

The common amino acids that contain asymmetric centres are not racemic, and all show, to varying degrees, an excess of the L-enantiomer^{7,8}. The $\delta^{15}\text{N}$ values could be determined for both enantiomers of alanine and glutamic acid. It is clear that these values are extraterrestrial (Table 1), and are essentially identical for the respective enantiomers of each amino acid within the present error of the measurement at the enrichments observed ($\pm 1\%$). The D/L value for alanine in this stone is 0.5 (Fig. 1), which is similar to that reported for alanine in the hydrolysed water extract of a previous stone (0.6; ref. 7). A similar value was obtained for yet a third stone (M.H.E. and J. A. Silfer, unpublished results). The D/L value for glutamic acid is consistent with our previously reported value of 0.3; ref. 7. Even if one assumes, for example, that the slight ^{15}N depletion in the L-alanine relative to D-alanine is real, it cannot be accounted for by a contribution from known terrestrial sources. If the original concentrations of the L-alanine and D-alanine had been identical, the excess L-alanine required to produce a D/L value of 0.5 would have to have an isotopic composition of approximately +55‰ to deplete L-alanine in this acid-hydrolysed water extract. To our knowledge, no terrestrial organic sources have been reported with this level of ^{15}N excess. Similarly, one can estimate the contribution of commonly observed terrestrial nitrogen (-10 to $+10\%$) to the isotope composition of the alanine, and its impact on the D/L value. This range of isotopic compositions would change the presumed initial value from 1.0 to 0.9. Thus it is clear that the $\delta^{15}\text{N}$ values reported here for Murchison amino acids are indigenous and have not been compromised by a terrestrial overprint.

An additional concern might be that the L-alanine or L-glutamic acid excess are the result of exotic co-eluting compounds. However, GC/MS and GC/C/IRMS analyses indicate no obvious co-elutions. Previous GC/MS analyses using a GC column with an entirely different stationary phase resulted in very similar L-enantiomer excesses⁷, also diminishing the likelihood of co-elutions. Moreover, if as yet unspecified compounds are co-eluting with the L-enantiomers, they would have to have similar structures (that would produce the same fragmentation ions on GC/MS), chemistry (for identical GC retention times) and isotopic compositions ($\delta^{15}\text{N}$ values) to enhance only the L-enantiomers. It should be noted that an L-enantiomer excess for several exotic amino acids has also recently been reported²⁸, although a stable-isotope assessment of their extraterrestrial origins was not attempted.

Table 1 Amino-acid abundances and $\delta^{15}\text{N}$ values

Amino acid	Concentration (nmol g ⁻¹)	$\delta^{15}\text{N}$ (‰)*
α -Aminoisobutyric acid	20.1	+184
Sarcosine	ND†	+129
Isovaline	8.0‡	+66
Glycine	24.5	+37
β -Alanine	12.8	+61
D-Alanine	—§	+60
L-Alanine	10.4‡	+57
L-Leucine	2.5§	+60
D,L-Proline	ND†	+50
D,L-Aspartic acid	4.7§	+61
D-Glutamic acid	—§	+60
L-Glutamic acid	10.8§	+58

* The $\delta^{15}\text{N}$ values are an average of four GC/C/IRMS analyses, the average error being about $\pm 1\%$. Values are reported relative to the standard, atmospheric N_2 : $\delta^{15}\text{N}(\text{‰}) = [(^{15}\text{N}/^{14}\text{N})_{\text{sample}} / (^{15}\text{N}/^{14}\text{N})_{\text{standard}} - 1] \times 10^3$.

† Not determined. Sarcosine was not sufficiently resolved by HPLC; the HPLC method did not detect secondary amines such as proline.

‡ The value for isovaline includes a contribution from valine that co-eluted with isovaline during HPLC analysis.

§ Concentrations reported for L-alanine, L-leucine, D,L-aspartic acid and L-glutamic acid represent the total contribution of both enantiomers for the respective amino acids.

|| $\delta^{15}\text{N}$ value reflects contribution of both enantiomers.

Our report of non-racemic amino acids is in conflict with the initial report of higher amino-acid D/L values for the Murchison meteorite¹⁰. However, we emphasize that our amino-acid distributions and relative abundances for all three stones are in general agreement with those initial and subsequent findings²⁶. This would not be expected if contamination occurred. It is also important to recognize that the amino acids present in the Murchison meteorite stones may not reflect the distribution at the initial time of synthesis. Organic and inorganic reactions may have altered amino-acid distributions over a period of unknown duration when the meteorite parent body (from which Murchison was derived) probably retained an aqueous phase⁹. Given this caveat, it is possible that different stones may contain components that do not reflect identical alteration histories. But although previous authors have attributed an apparently slight L-excess to biological contamination¹⁰, we believe, as will be discussed below, that the stable-isotope values point to an alternative explanation.

The moderate-to-extreme ^{15}N enrichments observed for amino acids in the Murchison meteorite establish their indigenous nature and, as previously suggested, point to a possible interstellar source for their precursors if not for some of the compounds themselves. Given the current hypotheses that (1) an enantiomeric excess is likely to have been a precondition for the origin of life on Earth (for example, refs 1, 2) and (2) the Earth's initial inventory of organic matter was derived from meteorite and comet bombardment⁶, the occurrence of an excess of the L-enantiomers in the Murchison meteorite provides the first evidence for a source of this chirality deemed essential for life's origin. The origins of the excess in the L-enantiomers presently remains unclear. The excess may have resulted from the alteration of initially abiotic, racemic mixtures by a process such as preferential decomposition by exposure to circularly polarized light¹. The extent to which such processes would have altered the stereochemistry of amino acids during the billion years preceding the first fossil evidence of life on Earth is unknown. But if amino-acid precursors (and perhaps some amino acids themselves) are probably interstellar⁹, this would certainly increase the time for exposure of organics to circularly polarized light before the formation of our Solar System. The preliminary report of a biotic signature in an Antarctic stone of martian origin²⁹, and the recent suggestion of life on Earth several hundred million years earlier than indicated by fossil evidence³⁰, certainly raises questions as to how pervasive life was in the early Solar System and the time required for its origin(s), in particular if chirality was a precondition for life. Stable-isotope analysis of amino-acid enantiomers, if present in the martian stone, may help to authenticate their origin(s)³¹, given the unique isotope signatures observed for Mars by the Viking mission³². □

Received 17 September 1996; accepted 14 July 1997.

- Bonner, W. A. The origin and amplification of biomolecular chirality. *Orig. Life* **21**, 59–111 (1991).
- Goldanskii, V. I. & Kuzmin, V. V. Chirality and cold origin of life. *Nature* **352**, 114 (1991).
- Cohen, J. Getting all turned around over the origins of life on Earth. *Science* **267**, 1265–1266 (1995).
- Oró, J. Comets and the formation of biochemical compounds on the primitive Earth. *Nature* **190**, 389–390 (1961).
- Chyba, C. F., Thomas, P. J., Brookshaw, L. & Sagan, C. Cometary delivery of organic molecules to the early Earth. *Science* **249**, 366–373 (1990).
- Chyba, C. F. & Sagan, C. Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life. *Nature* **355**, 125–132 (1992).
- Engel, M. H. & Nagy, B. Distribution and enantiomeric composition of amino acids in the Murchison meteorite. *Nature* **296**, 837–840 (1982).
- Engel, M. H., Macko, S. A. & Silfer, J. A. Carbon isotope composition of individual amino acids in the Murchison meteorite. *Nature* **348**, 47–49 (1990).
- Cronin, J. R. & Chang, S. in *The Chemistry of Life's Origins* (eds Greenberg, J. M. et al.) 209–258 (Kluwer Academic, Amsterdam, 1993).
- Kvenvolden, K. et al. Evidence for extraterrestrial amino acids and hydrocarbons in the Murchison meteorite. *Nature* **228**, 923–926 (1970).
- Epstein, S., Krishnamurthy, R. V., Cronin, J. R., Pizzarello, S. & Yuen, G. U. Unusual stable isotope ratios in amino acid and carboxylic acid extracts from the Murchison meteorite. *Nature* **326**, 477–479 (1987).
- Engel, M. H., Macko, S. A. & Nagy, B. in *Organic Geochemistry, Principles and Applications* (eds Engel, M. H. & Macko, S. A.) 685–695 (Plenum, New York, 1993).
- Hoppe, P., Amari, S., Zinner, E. & Lewis, R. S. Isotopic compositions of C, N, O, Mg and Si, trace element abundances, and morphologies of single circumstellar graphite grains in four density fractions from the Murchison meteorite. *Geochim. Cosmochim. Acta* **59**, 4029–4056 (1995).

14. Pizzarello, S., Krishnamurthy, R. V., Epstein, S. & Cronin, J. R. Isotopic analyses of amino acids from the Murchison meteorite. *Geochim. Cosmochim. Acta* **55**, 905–910 (1991).
15. Pizzarello, S., Feng, X., Epstein, S. & Cronin, J. R. Isotopic analyses of nitrogenous compounds from the Murchison meteorite: Ammonia, amines, amino acids, and polar hydrocarbons. *Geochim. Cosmochim. Acta* **58**, 5579–5587 (1994).
16. Fogel, M. L. & Cifuentes, L. A. in *Organic Geochemistry, Principles and Applications* (eds Engel, M. H. & Macko, S. A.) 73–98 (Plenum, New York, 1993).
17. Yuen, G., Blair, N., Des Marais, D. J. & Chang, S. Carbon isotopic composition of low molecular weight hydrocarbons and monocarboxylic acids from the Murchison meteorite. *Nature* **307**, 252–254 (1984).
18. Gilmour, I. & Pillinger, C. Isotopic differences between PAH isomers in Murchison. *Meteorites* **27**, 224–225 (1992).
19. Hare, P. E., St John, P. A. & Engel, M. H. in *Chemistry and Biochemistry of the Amino Acids* (ed. Barrett, G. C.) 415–425 (Chapman & Hall, London, 1985).
20. Merritt, D. A. & Hayes, J. M. Nitrogen isotopic analyses by isotope-ratio-monitoring gas chromatography/mass spectrometry. *J. Am. Soc. Mass Spectrom.* **5**, 387–397 (1994).
21. Macko, S. A., Uhle, M. E., Engel, M. H. & Andrushevich, V. Stable nitrogen isotope analysis of amino acid enantiomers by gas chromatography/combustion/isotope ratio mass spectrometry. *Anal. Chem.* **69**, 926–929 (1997).
22. Silfer, J. A. Stable carbon and nitrogen isotope signatures of amino acids as molecular probes in geologic systems. Thesis, Univ. Oklahoma (1991).
23. Silfer, J. A., Engel, M. H., Macko, S. A. & Jumeau, E. J. Stable carbon isotope analysis of amino acid enantiomers by conventional isotope ratio mass spectrometry and combined gas chromatography/isotope ratio mass spectrometry. *Anal. Chem.* **63**, 370–374 (1991).
24. Silfer, J. A., Qian, Y., Macko, S. A. & Engel, M. H. Stable carbon isotope compositions of individual amino acid enantiomers in mollusc shell by GC/C/IRMS. *Org. Geochem.* **21**, 603–609 (1994).
25. Grady, M. M., Wright, I. P., Swart, P. K. & Pillinger, C. T. The carbon and oxygen isotopic composition of meteoritic carbonates. *Geochim. Cosmochim. Acta* **52**, 2855–2866 (1988).
26. Shock, E. L. & Schulte, M. D. Summary and implications of reported amino acid concentrations in the Murchison meteorite. *Geochim. Cosmochim. Acta* **54**, 3159–3173 (1990).
27. Vallentyne, J. R. Biogeochemistry of organic matter-II. Thermal reaction kinetics and transformation products of amino compounds. *Geochim. Cosmochim. Acta* **28**, 157–188 (1964).
28. Cronin, J. R. & Pizzarello, S. Enantiomeric excesses in meteoritic amino acids. *Science* **275**, 951–955 (1997).
29. McKay, D. S. *et al.* Search for past life on Mars: Possible relic biogenic activity in Martian meteorite ALH84001. *Science* **273**, 924–930 (1996).
30. Mojzsis, S. J. *et al.* Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55–59 (1996).
31. Engel, M. H., Goodfriend, G. A., Qian, Y. & Macko, S. A. Indigeneity of organic matter in fossils: A test using stable isotope analysis of amino acid enantiomers in Quaternary mollusk shells. *Proc. Natl Acad. Sci. USA* **91**, 10475–10478 (1994).
32. Kieffer, H. H., Jakosky, B. M. & Snyder, C. W. in Mars (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 1–33 (Univ. Arizona Press, Tucson, 1992).

Acknowledgements. We thank S. Fulkerson for his support of this work, and V. Andrushevich and T. Brockwell for their assistance with the analyses. This work was supported by the US NSF. Instrument time and travel funds were provided by Micromass Inc.

Correspondence and requests for materials should be addressed to M.H.E. (e-mail: AB1635@OU.EDU).

Creating electrical contacts between metal particles using directed electrochemical growth

Jean-Claude Bradley, Hsuing-Ming Chen, Jeffrey Crawford, Jennifer Eckert, Karima Ernazarova, Thomas Kurzeja, Muduo Lin, Michael McGee, Wolfgang Nadler & Samuel G. Stephens

Drexel University, Department of Chemistry, 32nd and Chestnut Streets, Philadelphia, Pennsylvania 19104, USA

Electrical connections in microelectronics are usually established by means of photolithography to define the conducting channels. But methods that do not involve lithography have been explored, such as the use of electrodeposition¹ or electropolymerization^{2–6} to grow random structures of conducting material between two electrodes. This approach has been used to make diodes, transistors and signal amplifiers based on conducting polymers^{2,3}. Template-based^{7–12} and thermal plating¹³ strategies have also been used to direct the growth of electrically conducting media. One advantage of these approaches over photolithography is the possibility of forming contacts in three dimensions and so achieving enhanced data-processing densities. Previous electrochemical approaches have required that the electrodes to be connected are physically linked to the external voltage source. Here we show that electrodisolution and electrodeposition pro-

cesses in an applied electric field can be exploited to create directional growth of copper deposits between copper particles that are not connected to an external circuit. Moreover, the particles distort the electric field in such a way as to focus the diffusion of copper ions and consequently the direction of 'wire' growth, enabling the particles to be connected to one another in a directional and controllable manner. This suggests that appropriately directed electric fields may be used to connect an array of such particles into an arbitrary circuit pattern.

An electric field induces polarization in conductive particles. Beyond a critical polarization, the overpotential at the surface of the particle becomes sufficiently elevated to induce electrochemical reactions. Because each particle serves as both anode and cathode, the process is referred to as bipolar electrochemistry^{14–24}. This phenomenon has been investigated using fluidized or packed bed electrodes for applications in metal recovery^{14–17}, electrosynthesis^{18–21} and ultramicroelectrode^{22–24} studies. The technique is particularly well suited for electrochemistry in low-conductivity media. We use bipolar electrochemistry to form electrical contacts between metal particles physically isolated from an external circuit. This work is an extension of our previous interest in toposelective (site-selective) modifications of colloids using electric fields²⁵. In the present work, toposelective electrodisolution and electrodeposition are spatially coupled to generate copper deposits (referred to as 'wires') at predictable and highly selective locations.

The initial experimental conditions consist of aligning two copper particles perpendicular to two Pt electrodes in an aqueous medium (Figs 1 and 2). On application of an electric field, electrochemical oxidation of copper to cupric ions²⁶ occurs on one particle while electrochemical reduction of water occurs at the other particle. After a certain time the concentration of copper in the interparticle space increases to a level that allows electrodeposition to compete with the reduction of water. The result is the formation of fractal 'wires' which grow towards the nearest point on the other particle's surface. Once the wire spans the gap, electrical contact is achieved and all electrochemical phenomena between the particles instantly cease because the potential difference has been abolished. If one of the particles is placed near the cathodic feeder electrode, a wire will form between the particle and the electrode. This can easily be avoided by ensuring that the distance between the cathodic feeder electrode and the nearest particle is much larger than the interparticle distance.

The growth time and wire morphology can be readily controlled simply by changing the applied field intensity. At fields $< 15 \text{ V cm}^{-1}$ no wire formed within a period of 5 minutes (Fig. 3). A threshold field intensity for wire growth is expected as a minimum overpotential must be reached to ensure copper electro-oxidation and water reduction on the surface of the particles. At higher field intensities, the wire growth time rapidly decreased presumably owing to the accelerating release of copper ions and the increased electrophoretic force on these ions towards the particle acting as the cathode. Above about 35 V cm^{-1} the wire growth time was not significantly reduced with increasing field intensities. Such growth-speed independence on voltage has been observed in electrodeposition studies at very low metal ion concentrations²⁷. Typical non-bipolar electrodeposition experiments are governed by a complex interplay of diffusion, migration and electroconvection^{27–42}. It is thus surprising to find that the growth speed of the wire can be independent of voltage, even though the intense and inhomogeneous fields near the particles are almost certainly inducing powerful and complex convective flows.

As expected from the proposed mechanism, copper ions are initially generated by the anodic particle and must cross the interparticle gap. During this induction period no visible phenomena can be observed. We stress that this induction period is different from that observed in typical electrodeposition, which is thought to involve the building up of space charges or uniform

deposition^{28–31}. Wire growth then starts slowly and accelerates as the wire approaches the other particle where fields and copper concentrations are higher. Invariably, the wires had a fractal appearance, as expected for diffusion-limited aggregation^{27–42}. At higher potentials, the wires grew more quickly and appeared increasingly thin until an asymptotic speed was reached. Although difficult to resolve using light microscopy, the wire branches were estimated to be no more than a few micrometres in diameter.

During the initial growth period, several wires were frequently observed growing in parallel. An interesting positive-feedback mechanism seemed to take place, whereby the wire which was slightly ahead would grow faster than the others, effectively shielding the others from the field and presumably depleting the local copper-ion concentration. Competing with this process was the repetitive branching of the growing wire, generating a pseudo-steady-state of growing parallel branches generally connected to the same 'trunk' (Fig. 2c). Ultimately, the first branch to reach the other copper particle would close the circuit and all growth would cease instantly. As long as the field remained the wire showed self-healing behaviour: if the wire was broken at any point due to a slight

movement caused by tapping the microscope stage-plate, quick regrowth was observed unless the movement had caused complete detachment of the wire from both particles.

In the presence of externally added copper salts, wire growth was observed to occur not only between the particles but also from the particle facing the anodic feeder electrode: in addition, a thick bush was formed instead of a slender wire (compare Fig. 2g and i). Thus, the highly localized copper-ion distribution in the interparticle separation ensures that a wire will form exclusively between the particles. The ion localization had the added benefits that the overall conductivity of the solution could be kept low and electrodeposition of copper on the cathodic feeder electrode was minimized.

The distortion of the electric field by the two particles and the ensuing self-focusing behaviour could be observed by slanting the two particles relative to the external field (Fig. 2j). The field distortion by two or more conducting spheres and the ensuing increase in conductivity at the particle surface under bipolar conditions has recently been calculated numerically⁴³. However, the growth of fractal wires on the surface of one particle introduces significant complications in the calculation of the field distortion.

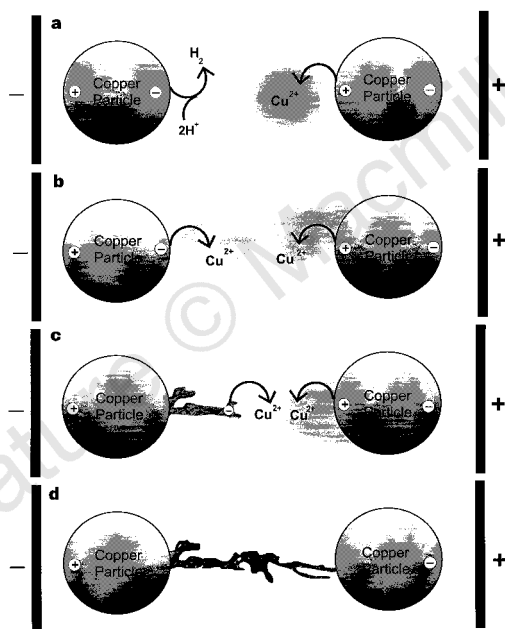


Figure 1 Diagram of wire formation between two particles under bipolar conditions. **a**, Two copper particles are placed in an aqueous environment and an electric field is applied. The polarization of each particle is shown. Initially the particle on the right liberates copper ions while the particle on the left reduces water. The shaded area represents a hypothetical distribution of the ionic cloud. (For clarity only the phenomena in the interparticle region are shown). **b**, When the copper-ion concentration near the particle on the left is high enough electro-deposition occurs and the wire begins to grow on the side facing the other particle. **c**, Electrodeposition occurs preferentially at the wire tip where cathodic polarization is expected to be highest. **d**, When the wire reaches the particle on the right, electrical contact is made. At this point there is no potential difference between the particles and electrochemical processes in the interparticle region cease.

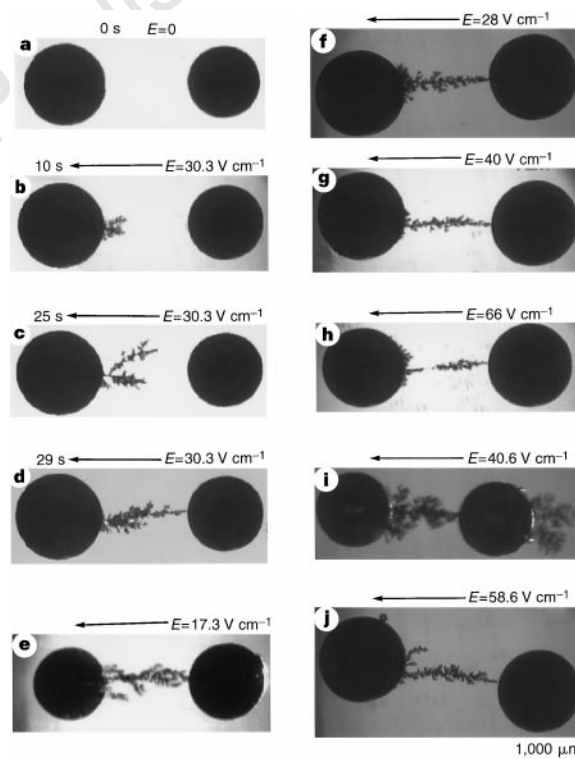


Figure 2 Wire formation between two copper particles. **a**, Before application of a 30.3 V cm⁻¹ field in the direction indicated by the black arrow. **b**, Several wires begin to grow on the left particle after a 10-s induction period. **c**, Two competing wire branches survive to the interparticle midpoint 25 s after the application of the field. **d**, Only one branch survives and spans the interparticle gap 29 s after the application of the field and establishes electrical contact between the two particles. We note the great acceleration in growth velocity as the wire approaches the second particle. **e–h**, Wires grown at increasing field strengths. Notice the reduction in branching and wire width with increasing fields. **i**, A wire grown between two particles with externally added 2.5 mM Cu(NO₃)₂. Instead of a thin wire forming exclusively between the particles, a thick bush forms both between the particles and on the particle closest to the feeder anode. **j**, A wire grown between two particles slightly slanted with respect to the external field showing that the shortest possible route was taken. Except for **i**, the aqueous medium contained 0.1 mM H₂SO₄ and 0.01% Nonidet-P40 (Sigma). The feeder electrodes were made of two parallel 1-mm-diameter Pt wires separated by 1.5 cm.

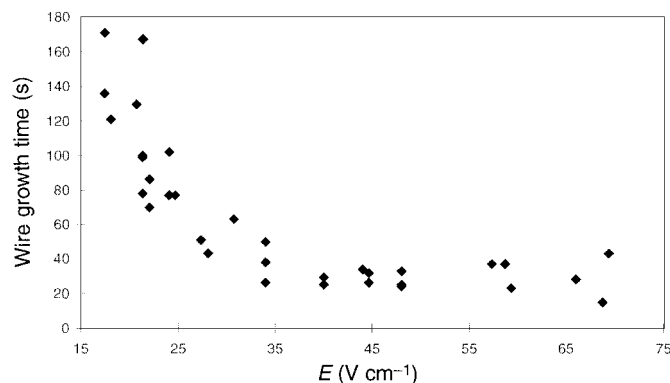


Figure 3 Dependence of wire growth time on the applied external field. Copper particles were $873 \pm 71 \mu\text{m}$ in diameter and the interparticle separation was $878 \pm 91 \mu\text{m}$. The same conditions as described in Fig. 2a–h were used. The wire growth time is defined as the interval between the application of the field and the formation of a wire connecting the two particles.

This process is further complicated by the highly inhomogeneous copper-ion distribution in the medium. We are attempting to model the kinetics and morphology of wire growth at various fields and interparticle distances.

As a first step towards the exploitation of bipolar electrochemistry to form microcircuits, we demonstrated the formation of wires in a particle array. Connections between selected particles could be achieved by simply changing the direction of the electric field vector (Fig. 4). By extending this concept to ordered three-dimensional particle arrays, we expect that the construction of three dimensional circuitry could be realized, which should permit far denser information processing than that available from two-dimensional lithographic techniques.

The ability to predict and control wire growth between metallic structures not directly connected to an external circuit may represent a simple and cost-effective approach to three-dimensional microcircuit building. For this to be feasible, it will be necessary to scale down the components and carry out the process in a suitable matrix. Because the polarization of a particle is directly proportional to its diameter, higher field intensities will be required for smaller structures. The high conductivity of aqueous systems may prove to be a significant limitation to achieving the required high fields, which might be overcome by using less polar solvents. Preliminary experiments have shown that submicrometre wires can be grown on particles as small as a few micrometres in diameter by a careful choice of solvent. Selective wire formation within a matrix-bound ensemble of conductive components might be achieved by using three-dimensional microelectrode arrays or perhaps even by the local electric field generated by linearly polarized light focused on a selected volume within the matrix. The use of a matrix may also prove to be an effective way of reducing the fragility of the wires. An especially powerful feature of this technique is the formation of wires much smaller than the metallic components initially present without having to resort to photolithographic methods. □

Received 30 April; accepted 7 July 1997.

- Gurtner, C. & Sailor, M. J. Selective construction of electrical connections using an organic charge transfer salt. *Adv. Mater.* **8**, 897–899 (1996).
- McCoy, C. H. & Wrighton, M. S. Potential dependent conductivity of conducting polymers yields opportunities for molecule based devices—a microelectrochemical push–pull amplifier based on 2 different conducting polymer transistors. *Chem. Mater.* **5**, 914–916 (1993).
- Jones, E. T. T., Chyan, O. M. & Wrighton, M. S. Preparation and characterisation of molecule-based transistors with a 50-nm source-drain separation with use of shadow deposition techniques—toward faster, more sensitive molecule-based devices. *J. Am. Chem. Soc.* **109**, 5526–5528 (1987).
- Sailor, M. J. & Curtis, C. L. Conducting polymer connections for molecular devices. *Adv. Mater.* **6**, 688–692 (1994).
- Fujii, M., Arai, K. & Yoshino, K. Neuron-type polypyrrole device prepared by electrochemical polymerization method and its properties. *Synth. Met.* **71**, 2223–2224 (1995).

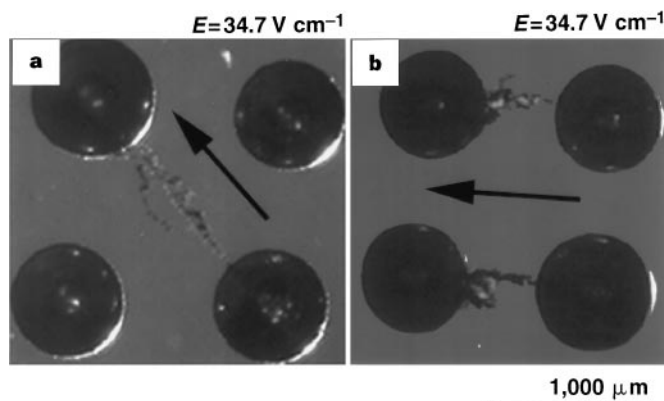


Figure 4 Control of wire growth within a 4×4 particle array. **a**, The applied electric field was along the diagonal of the copper particle array (indicated by the arrow) generating a single wire in the expected location. The other two particles were left unaffected. **b**, The applied electric field vector was parallel to the side of the particle array (arrow) generating two parallel wires.

- Curtis, C. L., Ritchie, J. E. & Sailor, M. J. Fabrication of conducting polymer interconnects. *Science* **262**, 2014–2016 (1993).
- Hidber, P. C., Nealey, P. F., Helbig, W. & Whitesides, G. M. New strategy for controlling the size and shape of metallic features formed by electroless deposition of copper—microcontact printing of catalysts on oriented polymers, followed by thermal shrinkage. *Langmuir* **12**, 4209–4215 (1996).
- Chou, S. Y., Krauss, P. R. & Renstrom, P. J. Imprint lithography with 25-nanometer resolution. *Science* **272**, 85–87 (1996).
- Wang, L., Yu-Zhang, K., Metrot, A., Bonhomme, P. & Troyon, M. TEM study of electrodeposited Ni/Cu multilayers in the form of nanowires. *Thin Solid Films* **288**, 86–89 (1996).
- Masuda, H. & Fukuda, K. Ordered metal nanohole arrays made by a 2-step replication of honeycomb structures of anodic alumina. *Science* **268**, 1466–1468 (1995).
- Martin, C. R. Nanomaterials—a membrane-based synthetic approach. *Science* **266**, 1961–1966 (1994).
- Huber, C. A. et al. Nanowire array composites. *Science* **263**, 800–802 (1994).
- Von Gutfield, R. J. & Vigliotti, D. R. Copper microcircuit repair of opens using thermally driven exchange plating. *Appl. Phys. Lett.* **56**, 2584–2586 (1990).
- Lee, J. K., Shemilt, L. W. & Chun, H. S. Studies of bipolarity in fluidized-bed electrodes. *J. Appl. Electrochem.* **19**, 877–881 (1989).
- Kusakabe, K., Kimura, T., Morooka, S. & Kato, Y. Recirculating feed operation of bipolar packed-bed and trickle-bed electrode cells equipped with mesh spacers. *J. Appl. Electrochem.* **17**, 724–730 (1987).
- Plimley, R. E. & Wright, A. R. A bipolar mechanism for charge transfer in a fluidized bed electrode. *Chem. Eng. Sci.* **39**, 395–405 (1984).
- Handley, D. & Eardley, D. C. Bipolar electrolysis with intraphase conduction in fluidized beds. *Chem. Ind.* 330–332 (1975).
- Sudoh, M., Kodera, T. & Ichino, T. Effect of oxygen sparging into packed-bed electrode on production rate of hydrogen peroxide. *J. Chem. Eng. Jpn.* **24**, 165–170 (1991).
- Manji, A. & Oloman, C. W. Electrosynthesis of propylene-oxide in a bipolar trickle-bed reactor. *J. Appl. Electrochem.* **17**, 532–544 (1987).
- Sudoh, M., Kodera, T., Hino, H. & Shimamura, H. Effect of anodic and cathodic reactions on oxidative-degradation of phenol in an undivided bipolar electrolyzer. *J. Chem. Eng. Jpn.* **21**, 198–208 (1988).
- Kim, H. J., Kusakabe, K., Hokazono, S., Morooka, S. & Kato, Y. Electro-oxidation rate of p-tertbutyltoluene in a bipolar packed bed electrode cell. *J. Appl. Electrochem.* **17**, 1213–1222 (1987).
- Ghoroghchian, J., Pons, S. & Fleischmann, M. Gas-phase electrochemistry on dispersions of microelectrodes. *J. Electroanal. Chem.* **317**, 101–108 (1991).
- Rolison, D. R. Zeolite-modified electrodes and electrode modified zeolites. *Chem. Rev.* **90**, 867–878 (1990).
- Fleischmann, M., Ghoroghchian, J., Rolison, D. & Pons, S. Electrochemical-behavior of dispersions of spherical ultramicroelectrodes. *J. Phys. Chem.* **90**, 6392–6400 (1986).
- Bradley, J.-C., Guedeau-Boudeville, M.-A., Jandau, G. & Lehn, J.-M. Toposomes bearing stable perforations by the action of an electric field on partially polymerized giant vesicles. *Langmuir* **13**, 2457–2462 (1997).
- Kirkova, L. & Rashkov, S. Anodic behavior of copper during electrorefining using a rotating ring-disc electrode. *Appl. Electrochem.* **24**, 420–425 (1994).
- Sawada, Y., Dougherty, A. & Gollub, J. P. Dendritic and fractal patterns in electrolytic metal depositions. *Phys. Rev. Lett.* **56**, 1260–1263 (1986).
- Fleury, V. & Barkley, D. Runaway growth in 2-dimensional electrodeposition. *Europhys. Lett.* **36**, 253–258 (1996).
- Bazant, M. Z. Regulation of ramified electrochemical growth by a diffusive wave. *Phys. Rev. E* **52**, 1903–1914 (1995).
- Cork, R. H., Pritchard, D. C. & Tam, W. Y. Local concentration measurements in electrochemical deposition using a schlieren method. *Phys. Rev. A* **44**, 6940–6943 (1991).
- Chazalviel, J. N. Electrochemical aspects of the generation of ramified metallic electrodeposits. *Phys. Rev. A* **42**, 7355–7367 (1990).
- Marshall, G. & Mocskos, P. Growth model for ramified electrochemical deposition in the presence of diffusion, migration, and electroconvection. *Phys. Rev. E* **55**, 549–563 (1997).
- Huang, W. G. & Hibbert, D. B. Fast fractal growth with diffusion, convection and migration by computer-simulation-effects of voltage of probability, morphology and fractal dimension of electrochemical growth in a rectangular cell. *Physica A* **233**, 888–898 (1996).
- Marshall, G., Tagtachian, S. & Lam, L. Growth-pattern formation in copper electrodeposition—experiments and computational modeling. *Chaos, Solitons Fractals* **6**, 325–339 (1995).
- Fleury, V., Kaufman, J. H. & Hibbert, D. B. Mechanism of a morphology transition in ramified electrochemical growth. *Nature* **367**, 435–438 (1994).

36. Barkey, D. P., Watt, D., Liu, Z. & Raber, S. The role of induced convection in branched electrodeposition morphology selection. *J. Electrochem. Soc.* **141**, 1206–1212 (1994).
37. Livermore, C. & Wong, P. Z. Convection and turbulence effects in strongly driven electrochemical deposition. *Phys. Rev. Lett.* **72**, 3847–3850 (1994).
38. Fleury, V., Chazalviel, J. N. & Rosso, M. Coupling of drift, diffusion, and electroconvection, in the vicinity of growing electrodeposits. *Phys. Rev. E* **48**, 1279–1295 (1993).
39. Halsey, T. C. Electrodeposition and diffusion-limited aggregation. *J. Chem. Phys.* **92**, 3756–3767 (1990).
40. Garik, P. *et al.* Laplace-field-controlled and diffusion-field-controlled growth in electrochemical deposition. *Phys. Rev. Lett.* **62**, 2703–2706 (1989).
41. Argoul, F., Arneodo, A., Grasseau, G. & Swinney, H. L. Self-similarity of diffusion-limited aggregates and electrodeposition clusters. *Phys. Rev. Lett.* **61**, 2558–2561 (1988).
42. Sawada, Y., Dougherty, A. & Gollub, J. P. Dendritic and fractal patterns in electrolytic metal deposits. *Phys. Rev. Lett.* **56**, 1260–1263 (1986).
43. Keh, H. J. & Li, W. J. Interactions among bipolar spheres in an electrolytic cell. *J. Electrochem. Soc.* **141**, 3103–3114 (1994).

Acknowledgements. We thank J. Yuan, M. Vallieres and A. Nath for discussions. This work was supported by Drexel University.

Correspondence and requests for materials should be addressed to J.-C.B. (e-mail: bradlecj@post.drexel.edu).

The interplay between binding energy and catalysis in the evolution of a catalytic antibody

Helle D. Ulrich^{††}, Emily Mundorff[†],
Bernard D. Santarsiero[†], Edward M. Driggers^{††},
Raymond C. Stevens[†] & Peter G. Schultz^{††}

^{*} Howard Hughes Medical Institute, and [†] Materials Science Division, Lawrence Berkeley National Laboratory and Department of Chemistry, University of California, Berkeley, Berkeley, California 94720, USA

Antibody catalysis¹ provides an opportunity to examine the evolution of binding energy and its relation to catalytic function in a system that has many parallels with natural enzymes. Here we report such a study involving an antibody AZ-28 that catalyses an oxy-Cope rearrangement, a pericyclic reaction that belongs to a well studied and widely used class of reactions in organic chemistry². Immunization with transition state analogue **1** results in a germline-encoded antibody that catalyses the rearrangement of hexadiene **2** to aldehyde **3** with a rate approaching that of a related pericyclic reaction catalysed by the enzyme chorismate mutase³. Affinity maturation gives antibody AZ-28, which has six amino acid substitutions, one of which results in a decrease in catalytic rate. To understand the relationship between binding and catalytic rate in this system we characterized a series of active-site mutants and determined the three-dimensional crystal structure of the complex of AZ-28 with the transition state analogue. This analysis indicates that the activation energy depends on a complex balance of several stereoelectronic effects which are controlled by an extensive network of binding interactions in the active site. Thus in this instance the combinatorial diversity of the immune system provided both an efficient catalyst for a reaction where no enzyme is known, as well as an opportunity to explore the mechanisms and evolution of biological catalysis.

Monoclonal antibody AZ-28 catalyses the unimolecular oxy-Cope rearrangement of substrate **2** to product **3** (which is trapped *in situ* as the oxime) with a rate acceleration ($k_{\text{cat}}/k_{\text{uncat}}$) of 5,300 (Fig. 1a)⁴. To determine the origins of the binding and catalytic properties of this antibody, its immunological precursor was cloned in the form of the germline variable-region (V_L and V_H, where L and H represent light and heavy chains, respectively) genes^{5,6} (Fig. 1b). During affinity maturation, the precursor to AZ-28 V_L, the T1/B V_L germline segment⁷, underwent five silent mutations as well as two amino-acid replacements: at position 34 on the light chain, serine was substituted by asparagine (Ser^{L34} to Asn) in complementarity-

determining region CDR1 and Ala^{L51} to Thr in CDR2. The precursor to AZ-28 V_H, the V_H 19.1.2 germline segment⁸, underwent four somatic mutations, Tyr^{H32} to Phe in CDR1, Ser^{H56} to Gly and Asn^{H58} to His in CDR2, and Thr^{H73} to Lys in framework region FR3, without any silent changes. CDR3 could be derived from the diversity (D) regions D_{SP2.2} (Asp^{H100}) or D_{SP2.3}, D_{SP2.4}, or D_{SP2.6} (Gly^{H100})⁹; alternatively, as AZ-28 was generated from an outbred mouse strain, it might be derived from an unmutated (wild-type) non-standard D region. Considering that a two-nucleotide change would be necessary to convert Gly or Asp to Phe^{H100}, this latter possibility seems more likely.

To determine the effects of affinity maturation on binding and catalysis, we reconstructed the pure germline (no somatic mutations) version of AZ-28. In addition, all six somatic mutations were reintroduced into the germline antibody individually as well as in combinations that varied according to their chains. Binding affinities and kinetics for hapten **1** were measured by surface plasmon resonance (Table 1). The heavy and light chains provide nine- and fivefold improvements in affinity, respectively, indicating that the effects of both chains are largely additive¹⁰. The resulting 40-fold overall decrease in the dissociation constant K_d (**1**) from 670 to 17 nM is well within the range reported for other antibodies to small haptens from secondary immune responses¹¹. Transition-state theory dictates that rate enhancement should correlate with increased differential binding of the antibody for the transition state relative to substrate^{12,13}. Surprisingly, however, the germline form of AZ-28, which has approximately the same Michaelis constant (K_M) and hence affinity for substrate as does mature AZ-28, affords a 35-fold greater rate enhancement compared with mature AZ-28 (the chemical step is rate-limiting in the case of the mature antibody)¹⁴. Moreover, mutation of Phe^{H100} to Asp in the germline antibody leads to an even greater rate acceleration ($k_{\text{cat}}/k_{\text{uncat}} = 3 \times 10^5$), despite a lower affinity for hapten **1**. This improvement in catalytic efficiency results in a rate enhancement approaching that of the well characterized enzyme chorismate mutase, which catalyses a related [3,3] sigmatropic rearrangement³.

Analysis of the germline single mutants indicates that whereas five of the six somatic mutations significantly affect binding affinity, the enhanced catalytic efficiency of the germline antibody compared to AZ-28 is associated largely with the mutation of Ser^{L34} to Asn. Kinetic data from the hybrid antibody L^{CH}wt (in which the light chain of the germline antibody is paired with the heavy chain of AZ-28) and the germline Ala^{L51}Thr mutant (Table 1) together indicate that the change of Asn^{L34} to Ser in the affinity matured antibody, AZ-28, leads to a 20-fold increase in k_{cat} . We further examined this dominant role of position L34 by site-directed mutagenesis. Substitution of Ala or Asp for Ser^{L34} in the germline antibody resulted in rate enhancements intermediate between that of the germline and mature antibody (Table 1), whereas substitution of His and Glu at this position led to decreases in activity of more than 250- and 2,500-fold, respectively. Measurement of the pH dependence of hapten dissociation rates in the range of pH 4.0 to 7.4 indicated that the only titratable group with an influence on hapten binding was Asp^{L34}, with the protonated form approaching the behaviour of Asn in this position. In contrast, the binding of the germline antibody and the germline mutants Ser^{L34} to Asn and Phe^{H100} to Asp, was essentially independent of pH¹⁵.

To understand the relationship better between the binding and catalytic activities of this series of related antibodies, we determined the three-dimensional crystal structure of the AZ-28 antigen-binding fragment (Fab) complexed with hapten **1** at 2.6 Å resolution. The hapten is bound in a deep, cylindrical cavity, about 8.3 Å wide and 18.5 Å deep; approximately 85% of hapten **1** (excluding linker) is buried within the Fab (Figs 2–4). Hapten **1** is bound in a chair-like geometry with the aryl and hydroxyl substituents of the cyclohexyl ring equatorial. The phenyl substituents of hapten **1** are rotated with respect to each other by a dihedral angle of 19° and make extensive

The X-ray crystal structure suggests that antibody AZ-28 functions by a combination of entropic¹⁶ and electronic effects. The antibody binds hapten **1** in a chair-like conformation, consistent with the preferred chair transition state of the Cope rearrangement¹⁷. The positions of the C2 and C5 atoms of hapten **1**, and as a consequence, the C2 and C5 atoms of hexadiene **2**, are fixed by extensive van der Waals interactions by the two phenyl substituents. Consequently, the bound conformation of the acyclic

Antibody	k_{cat} (min^{-1})	K_{M} (μM)	$k_{\text{cat}}/k_{\text{uncat}}$	k_{on} ($10^6 \text{ M}^{-1} \text{ s}^{-1}$)	k_{off} (10^{-2} s^{-1})	K_{d} (10^{-9} M)	$\Delta\Delta G$ (kcal mol^{-1})
AZ-28	0.023 (0.002)	74 (12)	5×10^3	3.11 (0.46)	0.55 (0.02)	17	-2.2
L ^{G1wt}	0.46 (0.03)	143 (14)	94×10^3	1.49 (0.08)	1.37 (0.02)	82	-1.2
H ^{G1wt}	0.076 (0.006)	205 (31)	16×10^3	0.47 (0.02)	0.64 (0.01)	152	-0.9
G	0.80 (0.06)	73 (13)	163×10^3	0.30 (0.01)	1.90 (0.08)	670	0.0
G(S ^{L34N})	0.076 (0.005)		16×10^3	0.49 (0.03)	0.70 (0.01)	150	-0.9
G(A ^{L51T})	0.83 (0.08)	78 (17)	169×10^3	0.31 (0.01)	1.59 (0.03)	686	0.0
G(S ^{L34A})	0.18 (0.01)	19 (3)	37×10^3	0.30 (0.02)	2.42 (0.04)	815	0.1
G(S ^{L34D})	0.22 (0.01)	108 (10)	45×10^3	0.19 (0.05)	2.96 (0.02)	1,282	0.4
G(Y ^{H32F})	0.66 (0.02)	59 (5)	135×10^3	1.04 (0.17)	1.62 (0.00)	162	-0.8
G(S ^{H56G})	0.75 (0.03)	78 (8)	153×10^3	0.77 (0.06)	1.79 (0.01)	206	-0.7
G(N ^{H58H})	0.71 (0.04)	90 (12)	145×10^3	0.50 (0.04)	1.21 (0.01)	211	-0.7
G(T ^{H73K})	0.81 (0.04)	75 (8)	165×10^3	0.55 (0.05)	1.62 (0.02)	309	-0.5
G(F ^{H100D})	1.45 (0.05)	123 (9)	296×10^3	0.39 (0.08)	2.75 (0.02)	840	0.1
G(F ^{H100G})	0.45 (0.01)	19 (1)	92×10^3	0.45 (0.03)	2.42 (0.01)	512	-0.2

Figure 1 a, Transition-state analogue (hapten **1, a** and **b** derivatives) and the oxy-Cope rearrangement reaction catalysed by antibody AZ-28. **b**, Light- and heavy-chain variable region sequences of AZ-28, and the germline precursors VL-T1/B and VH-19.12. Amino acids are represented in the one-letter code. Numbering is according to ref. 26. CDRs, complementarity determining regions, are indicated by bars.

NATURE | VOL 389 | 18 SEPTEMBER 1997

($4\pi + 2\sigma$) system of hexadiene **2** should be close to that of the transition state of either a stepwise or concerted rearrangement reaction^{18,19}. This is consistent with NMR studies that show transfer nuclear Overhauser effects (NOEs) between the terminal olefinic protons of substrate **2** when bound in the antibody active site¹⁴.

In addition to proper orbital alignment of the ($4\pi + 2\sigma$) orbitals, electronic effects arising from the 2,5-diphenyl and 3-hydroxyl substituents affect the energetics of 3,3-sigmatropic rearrangement reactions. Hyperconjugation of electron density on oxygen accelerates the oxy-Cope rearrangement through an anionic substituent effect²⁰. Consequently, His^{H56} and Glu^{H50} might act to enhance the rate of the rearrangement by increasing the electron density on the oxygen substituent of substrate **2**. The orientation of the latter residue may be affected by the somatic mutation of Asn^{H58} to His and Ser^{H56} to Gly. In addition to effects associated with the hydroxyl substituent, aryl substituents at the 2 and 5 positions of 1,5-hexadiene have also been shown to lower the activation energy by 5 to 10 kcal mol⁻¹ (this effect has been attributed to stabilization of a biradicaloid-like transition state)²¹. Hydrogen-bonding interactions with the hydroxyl group, however, probably rotate the two π bonds of the 1,5-hexadiene out of planarity with the two phenyl substituents, leading to decreased π -orbital overlap and an increase in

activation energy (Fig. 3). Thus the rate of the antibody AZ-28-catalysed reaction depends on a balance of the stereoelectronic effects involving alignment of the ($4\pi + 2\sigma$) orbitals of the diene, the degree of electron density on the hydroxyl substituent and the π -orbital overlap with the aryl substituents.

The crystal structure also provides an explanation for the inverse correlation of binding affinity to hapten **1** and catalytic rate. During the process of affinity maturation, increased binding interactions between antibody and hapten fix the orientation of the phenyl substituents with respect to the cyclohexyl ring. Consequently, upon substrate binding to mature AZ-28, the hexadiene core probably rotates with respect to the phenyl substituents to increase hydrogen-bonding and packing interactions (Fig. 3), making it more difficult to achieve maximal orbital overlap than in the case of the germline antibody, where the hapten is bound less tightly. Asn^{L34} is 3.8 Å from the cyclohexyl ring and directly interacts with Tyr^{H100a} and Asp^{H101} of V_HCDR3 to help fix the conformation of the bound hapten. Thus this residue might be expected to have the largest effect on the shape of the active site, and as a consequence on the stereoelectronic features of the reaction coordinate. Consistent with this model, variation in the structure of V_H CDR3, which is important for orienting the 2-phenyl substituent, also leads to large variations in

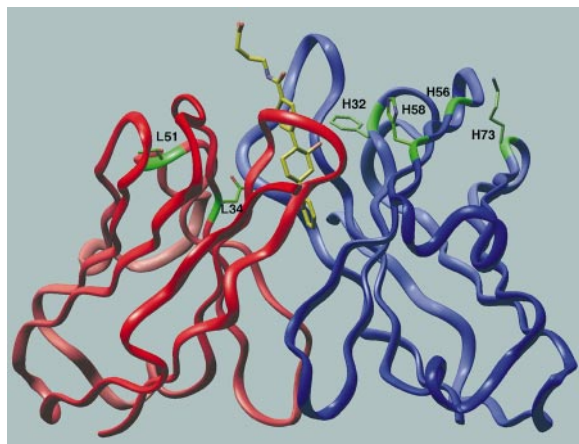


Figure 2 The three-dimensional structure of the variable domain of the AZ-28 Fab-hapten **1** complex showing the bound transition-state analogue (yellow) and sites of somatic mutation: Asn^{L34}, Thr^{L51}, Phe^{H32}, Gly^{H56}, His^{H58} and Lys^{H73} (green); the alpha-carbon (C α) backbone of V_H and V_L are in blue and red, respectively (heteroatom colours: O, red; N, blue).

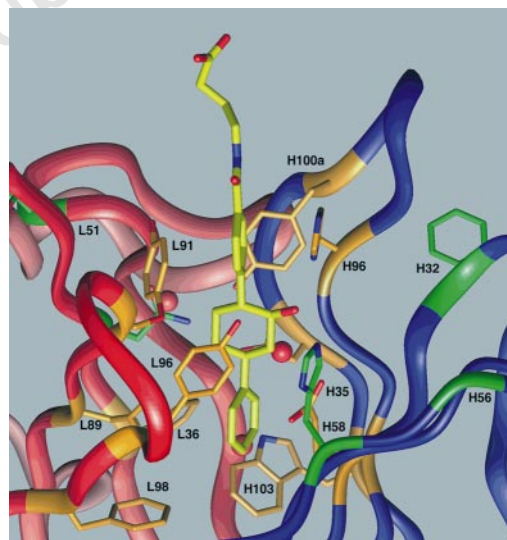


Figure 3 Structure of the AZ-28 active site showing the transition-state analogue **1** in yellow. The 5-phenyl ring is buried in the hydrophobic pocket and the 2-phenyl ring and linker are near the surface; the C α backbone of V_H and V_L are in blue and red, respectively, and sites of somatic mutation are indicated in green. The side chains of active-site residues interacting with hapten are indicated in brown. Two water molecules in the active site are indicated as red spheres (heteroatom colours: O, red; N, blue).

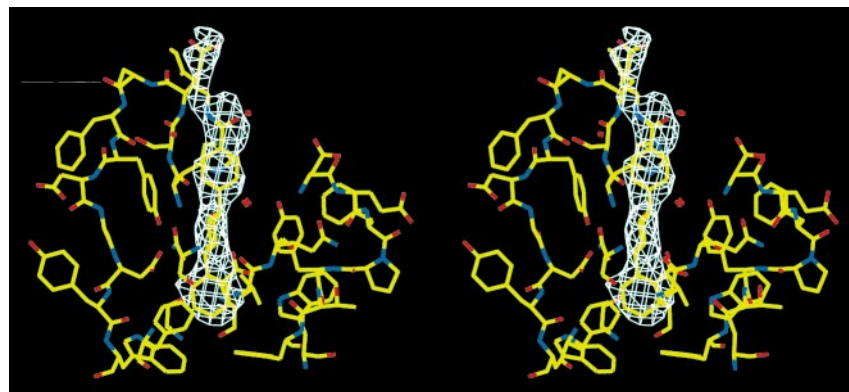


Figure 4 Stereodigram of a $1F_0 - F_c$ electron density map, illustrating the electron density in the antibody combining site. The electron density map was contoured at 2.2σ .

catalytic activity among a series of related antibodies, with V_H of AZ-28 being the most effective²².

Functional and structural analysis of the binding and catalytic properties of AZ-28 and related antibodies indicates that an extended network of hydrogen-bonding and hydrophobic interactions act together to control the stereoelectronic features of this reaction. Affinity based selection in response to hapten **1** leads to favourable orbital alignment of the (4 π + 2 σ) system of the diene and favourable interactions with the hydroxyl substituent. However, introduction of the somatic mutation Ser¹³⁴ to Asn, which leads to a modest increase in binding affinity, affects the relative orientation of the diene and phenyl orbitals, and therefore results in a significant decrease in rate²³. This inverse correlation of rate enhancement and affinity stems from the fact that rotation of the aryl substituents is not restricted in hapten **1**, as it is in the lowest-energy transition state. Nonetheless, the combinatorial diversity of the immune system, when selected by using mechanistic information programmed into hapten **1**, afforded an efficient antibody catalyst for an oxy-Cope rearrangement, a reaction for which there is no known enzyme. Improvements in hapten structure, together with mutations that optimize orbital alignment, should further increase catalytic efficiency^{24,25}. Finally, our study illustrates how catalytic antibodies can help us to understand the complex relationship between binding energy and catalysis in the evolution of biological catalysts. □

Methods

Cloning of germline antibody and construction of mutants. T1/B was identified as a potential germline candidate through a homology search of the Kabat database²⁶. Based on its published 5' flanking region⁷, a primer for the polymerase chain reaction (PCR) was designed to amplify the V_L-coding region and upstream flanking DNA from both AZ-28 hybridoma DNA and unrearranged murine liver DNA (of Swiss Webster origin). A primer based on the 3' untranslated region served for the amplification of unrearranged DNA, whereas primer J_k4 (ref. 5) was used for hybridoma DNA. Several clones of both PCR products were sequenced and their flanking regions were found to be identical with the published T1/B sequence. The PCR primer for analogous amplification of the V_H genes was designed based on the published 5' untranslated region of 37A11 V_H (ref. 27), a mutated version of 19.1.2 V_H (ref. 8). This primer, in combination with oligonucleotides covering the AZ-28 D region and the 3' end of the 19.1.2 V_H region, was used to amplify V_H and 319 nucleotides of 5' flanking region from AZ-28 hybridoma DNA and liver DNA, respectively. Two distinct 5' untranslated sequences were obtained from several clones of PCR product from liver DNA, one of which was identical to the corresponding region from AZ-28 hybridoma DNA. Its coding region—identical to that of 19.1.2—was therefore concluded to be the authentic AZ-28 germline sequence. The following modifications were made to the previously described expression system⁵: the M13 origin of pUC118 was inserted as blunt-ended *Apa*LI/*Nar*I fragment into the *Msc*I site of p4xH (ref. 5), yielding p4xH-M13. The V regions from pAZ-28 (ref. 6) were fitted by PCR with their authentic N termini, as determined from their germline sequences, and subcloned into p4xH-M13, resulting in pAZ-28^o. The germline V_H region was inserted in the same way, whereas the germline V_L region and all single mutants were constructed by site-directed mutagenesis, using sense oligonucleotides and avoiding rare codons.

Crystallization and structure determination. Antibody AZ-28 was crystallized in 25% PEG 1000, 0.1 M sodium acetate, pH 4.6, 0.3 M CdCl₂ and 0.1 M (NH₄)₂SO₄ in the presence of 2 mM hapten **1**, using the hanging-drop method. The crystals grew as very thin plates (0.4 mm × 0.3 mm × 0.02 mm). Data were collected at Stanford Synchrotron Radiation laboratory beamline 7-1 and processed using DENZO and SCALEPACK²⁸. The space group was P2₁, with two molecules in the asymmetric unit, and unit cell parameters were: *a* = 42.78 Å, *b* = 81.50 Å, *c* = 128.2 Å, β = 93.43 Å. The structure was solved using molecular replacement. The AMOR²⁹ package was used to search a library of variable domains using data between 8.0 and 3.5 Å. A solution for the variable domain was found with the model from 6FAB³⁰, with a correlation coefficient of 28.8% (*R* factor, 48.4%). Fixing this solution brought out the

second variable domain with a correlation coefficient of 28.6% (*R* factor, 48.9%). Mutation of the molecular replacement model and rebuilding was done using program O (ref. 31). Refinement was carried out using positional, simulated annealing and torsional refinement in X-PLOR, with NCS restraints turned on³². A total of 24,768 reflections were used in the refinement between 20.0–2.6 Å. Maps were produced using the CCP4 (ref. 33) package and interpreted in O. The electron density of the hapten was clearly observed in a 1F_o – F_c map. The data were 91.1% complete to 2.6 Å resolution with an *R*-merge of 6.0%. A total of 193 water molecules were placed in the structure. The final model has a free *R* of 30.0% and crystallographic *R* of 20.0%.

Characterization of the binding and catalytic properties of AZ-28 and related antibodies.

All antibodies were expressed and purified as recombinant Fab fragments in *Escherichia coli* as described^{5,6}. For surface plasmon resonance measurements, Fab fragments were additionally passed through a Superdex-75 column. Protein concentrations were determined spectrophotometrically based on calculated extinction coefficients at 280 nm. Structural integrity of the proteins was confirmed by circular dichroism. Catalytic constants were measured at 5 or 6 substrate concentrations in an HPLC assay⁴ in 20 mM MES, pH 6.0, 150 mM NaCl, 5 mM hydroxylamine, 4% acetonitrile, and 1% DMSO, with enantiomerically pure substrate⁶ and 30 μ M *o*-nitroanisole as internal standard. AZ-28 and H^oL^{wt} were used at 1.0 μ M, all other clones at 0.1 μ M. Rates were fitted to the Michaelis–Menten equation using the program Kaleidagraph. Affinity measurements were carried out in 20 mM MES, pH 6.0, 150 mM NaCl, and 0.005% P20 (Pharmacia) by surface plasmon resonance using a Biacore 2000 biosensor (Pharmacia). Sensor chips (CM5, research grade) were derivatized with the BSA-conjugate of hapten **1a** at 530–700 resonance units (RU) using the amine-coupling method. *K_d* values were determined under equilibrium conditions (20–30 min injections at 3 μ l min^{–1}, subtracting the signal obtained from a surface derivatized with unconjugated BSA) at 5 or 6 concentrations encompassing the *K_d*. Association rates were determined at 4 or 5 concentrations above the *K_d* at a flow rate of 20 μ l min^{–1}. The values of *k_{on}* were calculated from nonlinear curve fittings using the simple association model of the supplied software (Pharmacia Biosensor) and averaged. Fittings were started 10 s after the start of injection, covering 30–200 s of the association phase and avoiding phases of chip saturation. Residuals were in the range of <1 to a few RU. Dissociation rates were measured using saturation concentrations of Fab and a flow rate of 20 μ l min^{–1}. Fittings were from 10 to 40 s after the start of the dissociation phase. Increases in flow rate or inclusion of 400 μ M free hapten **1** had no further effect on the calculated *k_{off}* values (results not shown).

Received 15 April; accepted 21 July 1997.

- Schultz, P. G. & Lerner, R. A. From molecular diversity to catalysis: lessons from the immune system. *Science* **269**, 1835–1842 (1995).
- Woodward, R. B. & Hoffmann, R. *The Conservation of Orbital Symmetry* (Academic, New York, 1970).
- Andrews, P. R., Smith, G. D. & Young, I. G. Transition-state stabilization and enzymic catalysis. Kinetic and molecular orbital studies of the rearrangement of chorismate to prephenate. *Biochemistry* **12**, 3492–3498 (1973).
- Braisted, A. C. & Schultz, P. G. An antibody-catalyzed oxy-Cope rearrangement. *J. Am. Chem. Soc.* **116**, 2211–2212 (1994).
- Ulrich, H. D., Patten, P. A., Yang, P. L., Romesberg, F. E. & Schultz, P. G. Expression studies of catalytic antibodies. *Proc. Natl Acad. Sci. USA* **92**, 11907–11911 (1995).
- Ulrich, H. D., Driggers, E. M. & Schultz, P. G. Antibody catalysis of pericyclic reactions. *Acta Chem. Scand.* **50**, 328–332 (1996).
- Pech, M., Höchtl, J., Schell, H. & Zachau, H. G. Differences between germ-line and rearranged immunoglobulin V κ coding sequences suggest a localized mutation mechanism. *Nature* **291**, 668–670 (1981).
- Akolkar, P. N. *et al.* Different V_L and V_H germline genes are used to produce similar combining sites with specificity for α (1 → 6) dextran. *J. Immunol.* **138**, 4472–4470 (1987).
- Kurosawa, Y. & Tonegawa, S. Organization, structure, and assembly of immunoglobulin heavy chain diversity DNA segments. *J. Exp. Med.* **155**, 201–218 (1982).
- Patten, P. A. *et al.* The immunological evolution of catalysis. *Science* **271**, 1086–1091 (1996).
- Berek, C., Griffiths, G. M. & Milstein, C. Molecular events during maturation of the immune response to oxazolone. *Nature* **316**, 412–418 (1985).
- Haldane, J. B. S. *Enzymes* (Longmans, Green & Co., London, 1930).
- Pauling, L. *Chem. Eng. News* **24**, 1375 (1946).
- Driggers, E. M. *et al.* Mechanistic studies of an antibody-catalyzed oxy-Cope rearrangement. *J. Am. Chem. Soc.* (submitted).
- Ulrich, H. Recombinant catalytic antibodies. thesis, Univ. California, Berkeley, 1996.
- Page, M. I. & Jencks, W. P. Entropic contributions to rate accelerations in enzymic intramolecular reactions and the chelate effect. *Proc. Natl Acad. Sci. USA* **68**, 1678–1683 (1971).
- Doering, W. V. & Roth, W. R. The overlap of two allyl radicals or a four-centered transition state in the Cope rearrangement. *Tetrahedron* **18**, 67–74 (1962).
- Gajewski, J. J. Energy surfaces of sigmatropic shifts. *Acc. Chem. Res.* **13**, 142–148 (1980).
- Gajewski, J. J. & Gee, K. R. Solvent, counterion, and secondary deuterium kinetic isotope effects in the anionic oxy-Cope rearrangement. *J. Am. Chem. Soc.* **113**, 967–971 (1991).
- Steigewald, M. J., Goddard, W. A. & Evans, D. A. Theoretical studies of the oxy-anionic substituent effect. *J. Am. Chem. Soc.* **101**, 1994–1997 (1979).

21. Dewar, M. J. S. & Wade, L. E. A study of the mechanism of the cope rearrangement. *J. Am. Chem. Soc.* **99**, 4417–4424 (1977).
22. Ulrich, H. D. & Schultz, P. G. The evolution of binding affinity and catalysis in a family of catalytic antibodies. *J. Mol. Biol.* (submitted).
23. Storm, D. R. & Koshland, D. E. Jr A source for the special catalytic power of enzymes: Orbital steering. *Proc. Natl Acad. Sci. USA* **66**, 445–452 (1970).
24. Janda, K. D. et al. Direct selection for a catalytic mechanism from combinatorial antibody libraries. *Proc. Natl Acad. Sci. USA* **91**, 2532–2536 (1994).
25. Tawfik, D. S., Green, B. S., Chap, R., Sela, M. & Eshhar, Z. catELISA: A facile general route to catalytic antibodies. *Proc. Natl Acad. Sci. USA* **90**, 373–377 (1993).
26. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. *Sequences of Proteins of Immunological Interest* (US Dept of Health and Human Services, National Institutes of Health, Bethesda, 1991).
27. Nahmias, C., Strosberg, A. D. & Emorine, L. J. The immune response toward β -adrenergic ligands their receptors. VIII. Extensive diversity of V_H and V_L genes encoding anti-alprenolol antibodies. *J. Immunol.* **140**, 1304–1311 (1988).
28. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Meth. Enzymol.* **276**, 307–325 (1996).
29. Navaza, J. AMoRe: An automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
30. Strong, R. K. et al. Three-dimensional structure of a murine anti-P-azophenylarsenate FAB 36-71. 1. X-ray crystallography, site-directed mutagenesis, and modeling of the Complex with hapten. *Biochemistry* **30**, 3739–3748 (1991).
31. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
32. Brunger, A. X-PLOR, Version 3.851. A System for X-ray Crystallography and NMR (Yale Univ. Press, New Haven, Connecticut, 1996).
33. Collaborative Computational Project Number 4. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).

Acknowledgements. We thank M. Conn for precursor to the hapten, C. Cho and P. Yang for helpful discussions, and the Stanford Synchrotron Radiation Laboratory for synchrotron beam time. We are grateful for financial support for this work from the US Department of Energy and the NIH. P.G.S. is a HHMI investigator; H.D.U. was supported by a predoctoral fellowship from the NSF.

Correspondence and requests for materials should be addressed to R.C.S. or P.G.S.

Sources and contribution of terrigenous organic carbon to surface sediments in the Gulf of Mexico

Miguel A. Goñi[†], Kathleen C. Ruttenberg[†] & Timothy I. Eglinton[†]

^{*} Department of Geological Sciences, University of South Carolina, Columbia, South Carolina 29208, USA

[†] Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

The sources and burial processes of organic matter in marine sediments are not well understood, yet they are important if we are to have a better understanding of the global carbon cycle¹. In particular, the nature and fraction of the terrestrial organic carbon preserved in marine sediments is poorly constrained. Here we use the chemical and stable carbon isotope signatures of oxidation products from a macromolecular component (lignin)² of the terrigenous organic matter preserved in offshore surface sediments in the Gulf of Mexico to complement similar data from an existing onshore transect³ in this region. The complete onshore–offshore data set, along with radiocarbon dates of the bulk organic material at the same sites, allows the differentiation of material originating from plants that photosynthesize using the C₄ mechanism from those that undergo C₃ photosynthesis. We conclude that the offshore lignins derive from erosion of the extensive grassland (C₄) soils of the Mississippi River drainage basin, and that the nearshore lignins originate largely from C₃ plant detritus from coastal forests and swamps. This distribution is probably due to the hydrodynamic sorting of the different source materials⁴ during their seaward transport. These results suggest that previous studies^{3,5} have significantly underestimated the terrigenous fraction of organic matter in offshore sediments by not recognizing the contribution of C₄ vegetation to the carbon-isotope composition. Such an under-

estimate may force revisions in the assessment of past marine primary productivity and associated organic carbon fluxes⁶, and of organic matter preservation/remineralization⁷ and nutrient cycling⁸ in marine sediments.

The Gulf of Mexico is an extensively studied marine environment in which the fraction of organic carbon preserved in marine sediments that is of terrigenous origin (OC_{terr}) may be particularly high, owing to the proximity of continental land masses. Draining an area⁹ of 3.3×10^6 km², which includes most of North America's grasslands¹⁰ (Fig. 1), the Mississippi River system is the predominant source of sediment⁹ and fresh water¹¹ to the Gulf. We collected surface (0–2 cm) sediments from two transects perpendicular to the shore extending across the shelf and slope (Fig. 1). The elemental and isotopic compositions of organic matter from these sediments are shown in Table 1.

Conventional ¹⁴C ages of bulk sedimentary OC range between 2,580 and 6,770 yr BP, and generally increase with water depth. On the basis of published sedimentation rates¹², the average ages of marine phytodetritus in surface (0–2 cm) sediments should range between 3 (nearshore) and 400 (offshore) calendar yr. The large difference between ¹⁴C and estimated calendar ages suggests that, even after accounting for sediment mixing and the reservoir effect on marine OC (ref. 13), a significant fraction of the sedimentary OC must be relatively 'old', and most likely allochthonous in origin. Moreover, because OC_{terr} ¹⁴C ages are not subject to oceanic reservoir corrections, this material must be composed to a large extent of reworked OC (ref. 14). Non-zero apparent ¹⁴C ages for surface sediments are commonly observed on continental shelf or

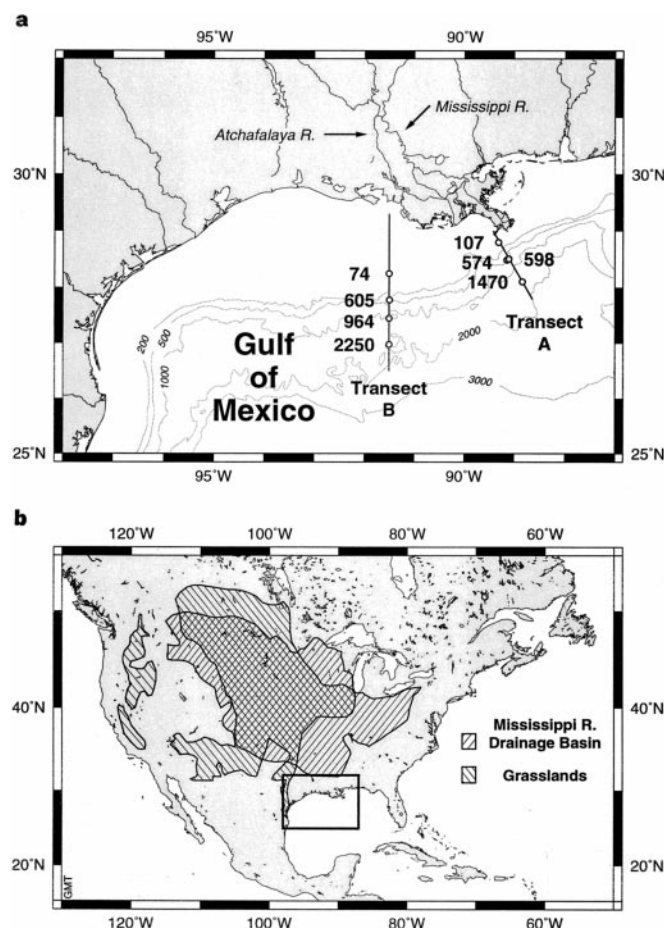


Figure 1 **a**, Map of the sampling locations and water depths (m) in the Gulf of Mexico. **b**, Map of the extent of the drainage basin of the Mississippi River and the grasslands in North America¹⁰. The area shown in **a** is an enlargement of the box in **b**.

slope environments, and are typically interpreted as resulting from relic OC_{terr} inputs¹⁵.

Sedimentary total organic carbon $\delta^{13}C$ values from both transects vary from -21.7 to -19.7‰ and display little trend with water depth¹². These generally enriched values agree well with previously published data (Fig. 2), and have been used as evidence for the dominance of autochthonous marine over terrestrially derived organic matter in offshore regions³. Such interpretations have assumed that the OC_{terr} reaching the Gulf of Mexico is derived primarily from C_3 vascular plants ($\delta^{13}C_{OC} \approx -26\text{‰}$). However, it is difficult to reconcile the proposed predominance of marine OC with the ^{14}C ages discussed above. Here we present data that contradict the assumption of an exclusively C_3 origin for OC_{terr} and thus may help resolve this apparent discrepancy.

Carbon-normalized yields of phenols derived from vascular (land) plant lignin are generally low (≤ 1.5 mg per 100 mg OC) for both transects (Table 1), and are consistent with previous observations (Fig. 2). Source parameters based on the phenol data from both transects show ratios of cinnamyl:vanillyl (C/V) ≥ 0.2 and syringyl:vanillyl (S/V) ≥ 1 (Table 1). These elevated C/V and S/V values indicate that lignin in these Gulf of Mexico sediments has a predominantly non-woody, angiosperm origin¹⁶ (for example, leaves and grasses). Because microbial degradation of vascular plant tissues alters both of these ratios¹⁷, it is imprudent to differentiate vascular plant sources much beyond this level of interpretation.

All sediments yield ratios of acid:aldehyde (ad/al) for both vanillyl and syringyl phenols that are ≥ 0.4 and generally increase offshore, reaching values as high as 3.5 (Table 1). Oxidative

Table 1 Compositional and isotopic parameters from Gulf of Mexico sediments

Parameter	Code	Depth of transect A (m)				Depth of transect B (m)			
		107	574	598	1,470	74	605	964	2,250
Organic carbon by weight (%)	% OC	1.4	1.4	1.3	0.9	0.9	1.2	0.6	0.3
Calcium carbonate by weight (%)	%CaCO ₃	2.0	3.8	3.8	16.6	7.7	11.5	28.7	25.4
Radiocarbon composition of organic carbon (‰)*	$\Delta^{14}C_{OC}$	-276.8	-291.1		-393.0	-367.0	-383.5	-386.4	-572.1
Radiocarbon age of organic carbon (yr BP)†	Age _{OC}	2,580	2,720		3,960	3,630	3,840	3,880	6,770
Lignin phenol yields (mg per 100 mg OC)									
Vanillyl phenols	V	0.563	0.309	0.309	0.085	0.202	0.159	0.136	0.143
Syringyl phenols	S	0.766	0.437	0.386	0.252	0.215	0.374	0.422	0.236
Cinnamyl phenols	C	0.156	0.085	0.107	0.070	0.081	0.072	0.092	0.102
Total	T _{lig}	1.49	0.83	0.80	0.41	0.50	0.60	0.65	0.48
Vanillyl + syringyl phenols‡	Λ	1.28	0.70	0.67	0.31	0.39	0.49	0.51	0.33
Lignin compositional parameters									
Syringyl:vanillyl ratio	S/V	1.36	1.41	1.25	2.98	1.07	2.35	3.09	1.65
Cinnamyl:vanillyl ratio	C/V	0.28	0.28	0.34	0.83	0.40	0.45	0.67	0.71
Vanillyl acid:aldehyde ratio	(ad/al) _V	0.56	0.76	0.67	0.46	0.75	1.16	1.18	0.78
Syringyl acid:aldehyde ratio	(ad/al) _S	0.41	1.00	0.53	1.51	1.53	0.46	0.61	3.47
Stable isotopic compositions (‰)§									
Bulk OC	$\delta^{13}C_{OC}$	-21.1		-20.4	-19.7	-20.9	-21.6	-20.2	-21.7
Syringaldehyde	$\delta^{13}C_{Si}$					-22.0			
Syringic acid	$\delta^{13}C_{Si}$	-29.0		-15.3	-15.2	-22.0	-21.8	-23.8	-19.9
p-Coumaric acid	$\delta^{13}C_{Cd}$	-28.5		-15.4					
Vanillin	$\delta^{13}C_{Vi}$	-26.7		-28.0		-29.9			

Where no data are given, $\delta^{13}C$ was not determined owing to unsatisfactory resolution by irm-GC-MS. Irm-GC-MS data for other lignin phenols not included in this table are not available because of their low overall and insufficient chromatographic resolution which precluded accurate $\delta^{13}C$ determination. In all samples, syringic acid was the most suitable of all lignin compounds for irm-GC-MS.

* Measured by AMS.

† The error in Age_{OC} ≤ 60 years for all samples.

‡ Sum of vanillyl + syringyl phenol yields (Λ) was calculated by excluding cinnamyl phenols to compare directly to previous data³.

§ All isotopic compositions, except for $\delta^{13}C_{OC}$ which was measured by direct combustion (precision of 0.1‰) were measured by isotope ratio monitoring-gas chromatography-mass spectrometry (irm-GC-MS).

|| Data from ref. 12. The precision of the irm-GC-MS measurements has been determined previously³, and for these samples is better than 1‰ for most compounds, but ranges up to 1.7‰ for syringic acid from the 605-m station.

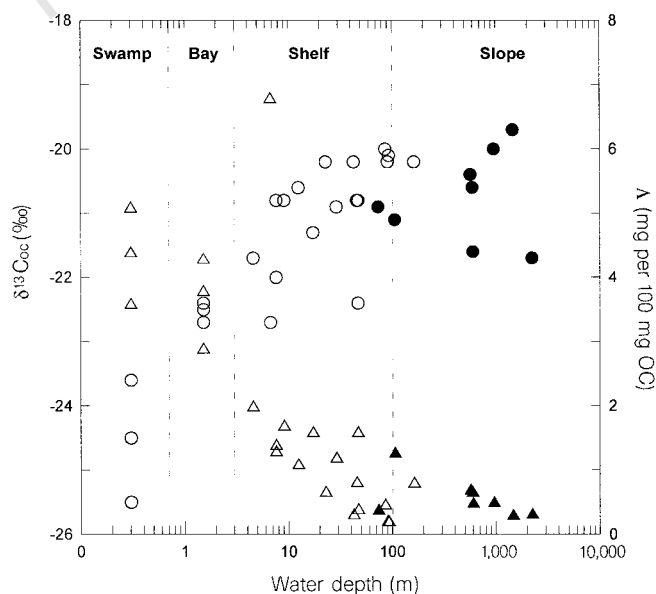


Figure 2 Distribution of $\delta^{13}C_{OC}$ (circles) and Λ (triangles) in surface sediments from various water depths of the Gulf of Mexico. Open symbols indicate data from several transects, including the Mississippi River south pass, southwest pass, Atchafalaya River and Terrebone Bay³. Filled symbols show data for samples from both transects (A and B) analysed in this study (Table 1).

degradation of lignin side-chains by microorganisms such as white-rot fungi often produces $\text{ad/al} > 0.4$ (ref. 18). Furthermore, materials containing highly altered lignin (such as fine suspended river sediments¹⁹ and humic horizons of mineral soils²⁰) display values ranging from 0.4 to 4. Hence the ad/al ratios reported here, which are consistent with previous data³, indicate that all sedimentary lignins have been subjected to a high degree of oxidative degradation.

In transect A, individual lignin phenols, most notably syringic acid, display significantly more depleted isotopic compositions at the 107-m site (-29‰) than those at deeper sites (-15‰). In contrast, the $\delta^{13}\text{C}$ values of most lignin phenols at all sites from transect B are relatively enriched ($\sim -22\text{‰}$). These compound-specific $\delta^{13}\text{C}$ values indicate that the sources of lignin to the Gulf of Mexico vary according to water depth and geographical location. Previously determined isotopic compositional ranges for individual lignin phenols² suggest that the OC_{terr} in the deeper sections of both transects has a dominant C_4 source. In contrast, OC_{terr} from the shallowest sections contains a high fraction of C_3 -derived lignin. Vanillin is different from other phenols because it is fairly depleted (-27 to -30‰) in sediments from both transects. One possible explanation for these vanillin compositions could be the occurrence of a small amount of C_3 lignin from gymnosperm wood, which would yield vanillyl phenols exclusively, in sedimentary mixtures otherwise containing lignin from non-woody angiosperm C_4 sources.

From the chemical (S/V, C/V and ad/al) and isotopic ($\delta^{13}\text{C}$ and $\Delta^{14}\text{C}$) data, we infer that the source of the OC_{terr} present in the slope and abyssal plain of this region of the Gulf is highly reworked, recalcitrant soil organic matter (probably sorbed to fine mineral grains) exported from the extensive grasslands of the Mississippi River drainage basin¹⁰. The vegetation in North American grasslands is composed of C_3 and C_4 plants, with cool-season C_3 species dominating in the northern ranges and warm-season C_4 species in the southern regions²¹. The $\delta^{13}\text{C}$ compositions of soil OC from the Great Plains correspond to the latitudinal distribution of vegetation and range from -14 to -23‰ (ref. 22). The ^{14}C -depleted values of OC and the highly degraded nature of lignins from offshore sediments are consistent with the bulk of OC_{terr} at these locations being composed of recalcitrant fractions of soil organic matter with turnover times exceeding 1,000 years (ref. 23). The low lignin yields (typical of recalcitrant soil organic-matter pools²⁰) obtained from the interior regions of the Gulf of Mexico do not therefore negate the possibility of a significant OC_{terr} contribution. In contrast, a large fraction of the OC_{terr} in the shelf regions of the Gulf of Mexico

(Fig. 2) seems to be less-degraded organic matter from C_3 sources, probably in the form of plant detritus from forests and swamps present in the lower reaches of the Mississippi River²⁴.

We do not yet know what causes these striking onshore-offshore trends in OC_{terr} composition. We propose that hydrodynamic sorting caused by the resuspension and cross-shelf transport of particles may serve to physically segregate distinct pools of OC_{terr} . The result of this process is the preferential transport of grassland soil organic matter closely associated with the finer-grained mineral load (clays) of the Mississippi River to interior regions of the northwestern Gulf of Mexico. In contrast, coarser and denser organic matter, containing waterlogged vascular plant detritus, is retained in bays, estuaries and the inner shelf, along with silt and sand-sized mineral particles. Such hydrodynamic sorting processes have previously been inferred to explain the chemical compositions of sedimentary organic matter off the Washington margin^{4,25} and the Amazon and Bengal deep sea fans^{26,27}. In this case, however, hydrodynamic sorting also seems to control the isotopic composition of OC_{terr} . Aeolian transport provides an additional means of contributing compositionally distinct OC_{terr} to the deeper regions of the Gulf of Mexico through entrainment of fine clay particles and their associated organic matter. However, the east-northeast trajectories of present-day prevailing winds suggest that this process is relatively unimportant.

Applying isotopic mass balance² to the $\delta^{13}\text{C}$ values of syringic acid, it is possible to evaluate the importance of C_4 lignin in sediments from both transects (Fig. 3). The relative abundances of lignin sources can then be used to estimate the average isotopic signature of OC_{terr} (Fig. 3). These calculations, bearing in mind the assumptions involved², show that C_4 -derived OC modifies the isotopic signature of OC_{terr} ($\delta^{13}\text{C}_{\text{terr}} = -20\text{‰}$ to -22‰) in offshore Gulf sediments such that it is indistinguishable from marine OC ($\delta^{13}\text{C}_{\text{mar}} = -19\text{‰}$ to -21‰). In such instances, quantitative assessments of terrigenous organic matter inputs using bulk $\delta^{13}\text{C}_{\text{OC}}$ are difficult to accomplish, and may have led to an underestimation of allochthonous contributions^{3,5}.

Grasslands (both tropical and temperate) presently occupy nearly one-fifth of the Earth's surface²⁸. Given their wide distribution, our findings underscore the potential importance of grassland soils as sources of OC_{terr} to present-day oceans. The distribution and composition of grasslands are intimately connected with climate^{22,28}, so, climatic variability may have affected the export of OC_{terr} not only to contemporary but also to past oceans. For example, increased input of terrigenous C_3 carbon to the Gulf of Mexico during periods of low sea level has been inferred from the

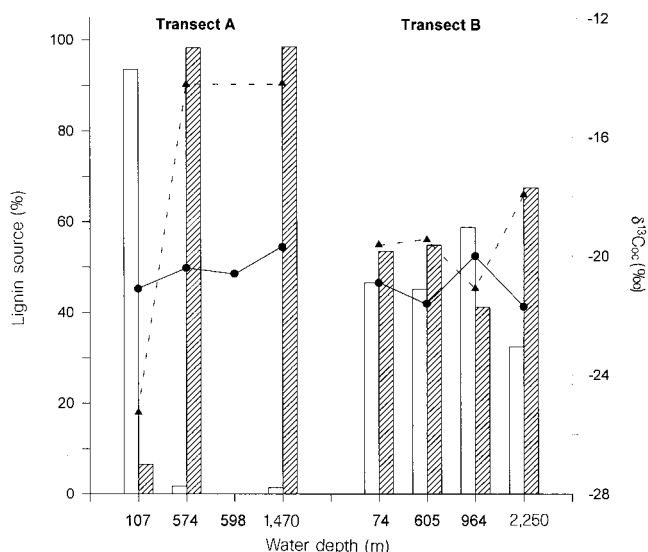


Figure 3 Relative contribution of C_3 (white bars) and C_4 (shaded bars) vascular plants as sources of sedimentary lignin in the Gulf of Mexico, based on the isotopic composition of syringic acid (measured by irm-GC-MS) from sediments (Table 1) and from C_3 and C_4 end-members² (-30‰ and -15‰ , respectively). The estimated isotopic composition of terrigenous organic carbon ($\delta^{13}\text{C}_{\text{terr}}$, triangles) and the measured $\delta^{13}\text{C}$ values bulk organic carbon ($\delta^{13}\text{C}_{\text{OC}}$, circles) are also shown. Details of the approach used to estimate $\delta^{13}\text{C}_{\text{terr}}$ have been discussed previously². Briefly, using isotopic mass balance, it is possible to estimate the relative abundance of C_3 and C_4 terrestrial organic matter sources in marine sediments from the $\delta^{13}\text{C}$ values of individual lignin phenols. However, the same lignin phenol may be derived from two different sources in vastly different quantities, depending on the nature and composition of the precursor (such as plant tissues versus soil OC). Hence, when interpreting these data, the role of oxidation yield in determining the isotopic signature of a given CuO product must be considered². Given these constraints, we used the $\delta^{13}\text{C}$ values of the best-resolved lignin phenol, syringic acid (Table 1), to semi-quantitatively estimate the fraction of C_3 and C_4 angiosperm lignin in Gulf of Mexico sediments. This choice appears justified because the lignin in most of the sediments analysed in this study seems to be derived from highly degraded angiosperm sources, and syringic acid is the best marker available for such material.

considerably depleted $\delta^{13}\text{C}_{\text{OC}}$ signatures obtained from sediments deposited during glacial epochs⁵. However, the observed glacial–interglacial isotopic variations could instead reflect changes from a C_3 -dominated (colder temperature) vegetation in the Mississippi River drainage basin during glaciations to a more C_4 -dominated vegetation in the Holocene epoch. This latter interpretation is supported by the known temperature-dependent latitudinal distribution of plants in present-day North American grasslands²¹. Application of these and other biological marker $\delta^{13}\text{C}$ data²⁹, in combination with compound-specific radiocarbon analysis¹⁴, to downcore studies of marine sediments may help us to understand the changes in vegetation (C_3 versus C_4) resulting from continental climate change in North America and other areas of the world^{22,27–30}. □

Received 11 July; accepted 1 August 1997.

- Hedges, J. I. Global biogeochemical cycles: Progress and problems. *Mar. Chem.* **39**, 67–93 (1992).
- Goñi, M. A. & Eglinton, T. I. Stable carbon isotopic analyses of lignin-derived CuO oxidation products by isotope ratio monitoring–Gas Chromatography–Mass spectrometry (irm–GC–MS). *Org. Geochem.* **24**, 601–615 (1996).
- Hedges, J. I. & Parker, P. L. Land-derived organic matter in surface sediments from the Gulf of Mexico. *Geochim. Cosmochim. Acta* **40**, 1019–1029 (1976).
- Keil, R. G., Tsamakis, E., Bor Fuh, C., Giddings, J. C. & Hedges, J. I. Mineralogical and textural controls on the organic composition of coastal marine sediments: hydrodynamic separation using SPLITT-fractionation. *Geochim. Cosmochim. Acta* **58**, 879–893 (1994).
- Jasper, J. P. & Gagosian, R. B. Glacial–interglacial climatically forced $\delta^{13}\text{C}$ variations in sedimentary organic matter. *Nature* **342**, 60–62 (1989).
- Müller, P. J. & Suess, E. Productivity, sedimentation rate, and sedimentary organic matter in the ocean. I. Organic carbon preservation. *Deep Sea Res.* **26**, 1347–1367 (1979).
- Hedges, J. I. & Keil, R. G. Sedimentary organic matter preservation: an assessment and speculative synthesis. *Mar. Chem.* **49**, 81–115 (1995).
- Martin, J. H., Knauer, G. A., Karl, D. M. & Broenkow, W. W. VERTEX: Carbon cycling in the northeast Pacific. *Deep Sea Res.* **34**, 267–285 (1987).
- Van Andel, T. H. in *Recent Sediments, Northwestern Gulf of Mexico* (eds Shepard, F. P., Phleger, F. B. & Van Andel, T. H.) 34–55 (American Assoc. of Petrology and Geology, Tulsa, 1960).
- Sims, P. L. & Coupland, R. T. Producers in *Grassland Ecosystems of the World: Analyses of Grasslands and their Uses* (ed. Coupland, R. T.) 49–72 (Cambridge Univ. Press, 1979).
- Deegan, L. A. *et al.* in *Estuary Variability* (ed. Wolfe, D. A.) 83–100 (Academic, New York, 1986).
- Ruttenberg, K. C. & Goñi, M. A. Phosphorus distribution, C:N:P ratios, and $\delta^{13}\text{C}_{\text{OC}}$ in arctic, temperate, and tropical coastal sediments: tools for characterizing bulk sedimentary organic matter. *Mar. Geol.* **139**, 123–145 (1997).
- Stuiver, M., Pearson, G. W. & Braziunas, T. Radiocarbon age calibration of marine samples back to 9,000 cal yr. bp. *Radiocarbon* **28**, 980–1021 (1986).
- Eglinton, T. I. *et al.* Variability in radiocarbon ages of individual organic compounds from marine sediments. *Science* **277**, 796–799 (1997).
- Emerson, S. C. *et al.* Organic carbon in surface deep-sea sediments: C-14 concentration. *Nature* **329**, 51–54 (1987).
- Hedges, J. I. & Mann, D. C. The characterization of plant tissues by their cupric oxide oxidation products. *Geochim. Cosmochim. Acta* **43**, 1803–1807 (1979).
- Opsahl, S. & Benner, R. Early diagenesis of vascular plant tissues: lignin and cutin decomposition and biogeochemical implications. *Geochim. Cosmochim. Acta* **59**, 4889–4904 (1995).
- Goñi, M. A., Nelson, B., Blanchette, R. A. & Hedges, J. H. Fungal degradation of wood lignins: geochemical perspectives from CuO -derived phenolic dimers and monomers. *Geochim. Cosmochim. Acta* **57**, 3985–4002 (1993).
- Hedges, J. I. *et al.* Compositions and fluxes of particulate organic material in the Amazon river. *Limnol. Oceanogr.* **31**, 717–738 (1986).
- Ziegler, F., Kögel, I. & Zech, W. Z. Alteration of gymnosperm and angiosperm lignin during decomposition in forest humus layers. *Z. Pflanzenenergie. Bodenkr.* **149**, 323–331 (1986).
- Teeri, J. A. & Stowe, L. G. Climatic patterns and the distribution of C_4 grasses in North America. *Oecologia* **23**, 1–12 (1976).
- Tieszen, L. L. & Archer, S. in *Plant Biology of the Basin and Range* (eds Osmond, C. S., Pitelka, L. F. & Hidy, G. M.) 293–321 (Ecological Studies Ser. Vol. 80, Springer, Heidelberg, 1990).
- Parton, W. J., Schimel, D. S., Cole, C. V. & Ojima, D. S. Analysis of factors controlling soil organic matter levels in Great Plains grasslands. *Soil Sci. Soc. Am. J.* **51**, 1173–1179 (1987).
- Chmura, G. L. Palynomorph distribution in marsh environments in the modern Mississippi Delta Plain. *Geol. Soc. Am. Bull.* **106**, 705–714 (1994).
- Prahl, F. G., Ertel, J. R., Goñi, M. A., Sparrow, M. A. & Eversmeyer, B. Terrestrial organic carbon contributions to sediments on the Washington margin. *Geochim. Cosmochim. Acta* **58**, 3035–3048 (1994).
- Goñi, M. A. in *Proceedings of the Ocean Drilling Program, Sci. Rep. 155* (eds Flood, R. D., Piper, D. J. W. & Klaus, A.) 519–530 (College Station, TX, 1997).
- France-Lanord, C. & Derry, L. A. $\delta^{13}\text{C}$ of organic carbon in the Bengal Fan: source evolution and transport of C_3 and C_4 plant carbon to marine sediments. *Geochim. Cosmochim. Acta* **58**, 4809–4814 (1994).
- Parton, W. J. *et al.* Observations and modeling of biomass and soil organic matter dynamics for the grassland biome worldwide. *Global Biochem. Cycles* **7**, 785–809 (1993).
- Bird, M. I. *et al.* Terrestrial vegetation change inferred from n -alkane $\delta^{13}\text{C}$ analysis in the marine environment. *Geochim. Cosmochim. Acta* **59**, 2853–2857 (1995).
- Cerling, T. E., Wang, Y. & Quade, J. Expansion of C_4 ecosystems as an indicator of global ecological change in the late Miocene. *Nature* **361**, 344–345 (1993).

Acknowledgements. We thank J. W. Morse for the opportunity to participate on the cruise in Gulf of Mexico; E. Tsamakis, K. Tholke, R. Ostermann, J. Irvine, N. Parmentier, L. Christman, J. Donahue and E. Tappa for help with sample preparation and analyses; and G. Eglinton, B. Fry, J. Jasper, J. Hedges, R. Benner, S. Smith and S. Montgomerie for comments and discussions. This work was supported by the US NSF.

Correspondence and requests for materials should be addressed to M.A.G. (e-mail: goni@epoch.geol.sc.edu).

Trade-off between parasitoid resistance and larval competitive ability in *Drosophila melanogaster*

A. R. Kraaijeveld & H. C. J. Godfray

NERC Centre for Population Biology and Department of Biology, Imperial College at Silwood Park, Ascot, Berkshire SL5 7PY, UK

The extent to which an organism is selected to invest in defences against pathogens and parasites depends on the advantages that ensue should infection occur, but also on the costs of maintaining defences in the absence of infection. The presence of heritable variation in resistance suggests that costs exist, but we know very little about the nature or magnitude of these costs in natural populations of animals¹. A powerful technique for identifying trade-offs between fitness components is the study of correlated responses to artificial selection^{2,3}. We have selected *Drosophila melanogaster* for improved resistance against an endoparasitoid, *Asobara tabida*. Endoparasitoids are insects whose larvae develop internally within the body of other insects, eventually killing them, although their hosts can sometimes survive attack by mounting a cellular immune response^{4–6}. We found that reduced larval competitive ability in unparasitized *D. melanogaster* is a correlated response to artificial selection for improved resistance against *A. tabida*. The strength of selection for competitive ability and parasitoid resistance is likely to vary temporally and spatially, which may explain the observed heritable variation in resistance.

Parasitoids are very common natural enemies of many insects, developing either externally or internally on their hosts⁷. Endoparasitoids frequently suspend development as eggs or first-instar larvae while their hosts feed and grow in size, although during this period they can be destroyed by encapsulation, a cellular immune response. Parasitoids avoid host defences by hiding in tissue away from circulating haemocytes, by avoidance of recognition through molecular mimicry, or through debilitating the immune response by toxins or symbiotic viruses⁸. Individual hosts often differ in their ability to encapsulate parasitoids, and selection, quantitative genetic and isofemale-line studies have shown this to have a genetic basis^{8–11}. The presence of additive genetic variation in encapsulation ability suggests that there are costs to the possession of defences against parasitoids, although these costs have yet to be identified in any host–parasitoid interaction.

Asobara tabida (Hymenoptera, Braconidae) is a common larval parasitoid of *Drosophila* species on fermenting substrates in Europe. In central and northern Europe its main host is *D. subobscura*, in which it is never encapsulated, although it is also found on *D. melanogaster* at low frequencies¹². In Mediterranean Europe *D. melanogaster* is the main host. Our base stock population of *D. melanogaster* was derived from 250 flies caught in the wild near Leiden in the Netherlands and maintained in the laboratory for three generations before the start of the experiment. The wasps used in the selection experiments were a laboratory strain derived originally from insects caught in Sospel, France, and cultured on *D. subobscura* to prevent adaptation to *D. melanogaster*. The base stock encapsulated only 5% of the eggs of the laboratory strain of *A. tabida*, a figure typical of fly populations from northern Europe¹³.

The base stock was split into eight lines, four of which were selected for increased encapsulation ability, the others acting as controls. Each new generation of the selection lines was bred from flies that had survived parasitoid attack (the encapsulated parasitoid egg remains in the body of the fly and can be seen through the wall of

the puparium). To minimize inbreeding, a population size of at least 100 was maintained in all lines. With the exception of exposure to parasitoids, control and selection lines were cultured using identical procedures. Larvae were exposed to parasitoids in the selection treatment in conditions that lead to high rates of parasitism to avoid selection on traits other than parasitoid resistance. The encapsulation ability of the flies was assessed by exposing 200 second-instar larvae to parasitoid females for 2 h, after which the larvae were kept at 20 °C for five days and then dissected to score capsule frequency.

The response to selection was rapid and similar across the four lines (Fig. 1). The probability of encapsulation rose from 5% to a maximum of ~60% in five generations. The maximum encapsulation ability is similar to that found in Mediterranean populations, which are heavily attacked by *A. tabida*¹³. After selection, four pairs of F₁ lines were created by crossing each of the selected lines with a different control line. In one cross of the pair, males derived from selected lines and females from control lines, whereas in the other cross the sexes were reversed. The mean encapsulation rate of the offspring was 34% (standard error 4%), and was not influenced by the origin of the female parent (*t*-test, *P* = 0.9), excluding maternal or cytoplasmic effects, and strongly suggesting that the genes influencing resistance are not sex-linked.

We searched for correlated responses to selection by comparing a range of fitness parameters in the control and selected lines. In a first set of experiments we compared the performance of flies reared under conditions of abundant food and very low competition. The full details of the experiments will be reported elsewhere, but in summary we found no significant differences in larval and pupal survival, larval and pupal development time, adult longevity in the absence of food, early female fecundity, adult size or fluctuating asymmetry. Working with only eight lines in total we have limited statistical power to detect subtle differences, but in no cases were strong but insignificant trends detected.

In a second set of experiments we compared the performance of control and selected flies under conditions of weak to strong intraspecific competition (Fig. 2). Competitive ability was assessed by rearing larvae with flies from a genetically marked 'tester' stock^{14,15} at four different levels of competition for food (see

Methods for experimental and statistical details). When competition was weak (0.4 ml and 0.2 ml larval food), survival was high (~80%) and there was no difference between the two sets of lines (*P* > 0.1). However, for more severe competition (0.1 ml), survival was reduced and there was significantly greater mortality among selected than control flies (*P* = 0.011). This difference was also seen at the most severe level of competition (0.05 ml), where survival was much lower (~50%), although the statistical significance was less (*P* = 0.061) owing to the much greater within-line variability in survival when resources were very scarce. At higher levels of competition, flies are smaller, take longer to develop, and show more fluctuating asymmetry, but we found no significant differences between the performances of the control and selected lines (although there was a trend for selected flies to be smaller at the two lower levels of competition).

Thus reduced survival in a high-competition environment is a correlated response to selection for improved defence against endoparasitoid attack in *D. melanogaster*. This suggests there is a trade-off between defence against parasitoids and other components of fitness. We suspect that selected larvae allocate more resources to the machinery of cellular encapsulation or to counter-acting the wasp's attempts to disable encapsulation, and hence are less able to withstand very stressful conditions. Several other trade-offs in the life-history strategies of *D. melanogaster* have been identified only in circumstances when the fly is under stress². *D. melanogaster* populations in the wild feed in rotting fruit and other fermenting substrates, and often experience a level of competition in the field comparable with those at which survival of the selected larvae was reduced¹⁶.

We believe this to be the first demonstration of a cost of resistance against parasitoids. For many species of hosts, parasitoid attack rates vary both temporally and spatially, and the resulting fluctuating selection pressures, combined with costs to parasitoid defences, may explain the additive genetic variance in defensive ability observed in several host species^{8–10}. The population dynamics of hosts and their parasitoids are often coupled, and our findings suggest that there may be complex joint evolutionary and population dynamic interactions. More generally, evidence for costs of

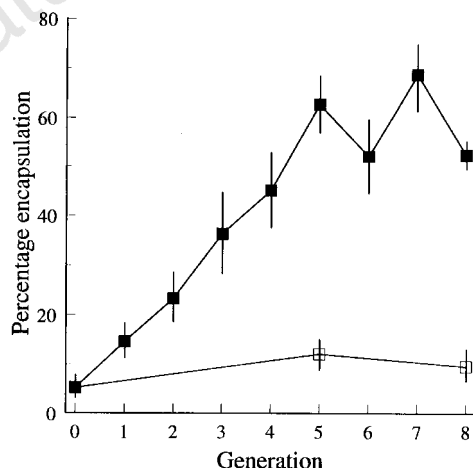


Figure 1 The frequency of encapsulation in control and selected flies, showing means and standard errors of the four selected and control lines. For logistical reasons (assessment of encapsulation is very time consuming), control lines were not assessed every generation. Differences were significant in generations 5 (*t*-test, *P* < 0.01) and 8 (*t*-test, *P* < 0.01).

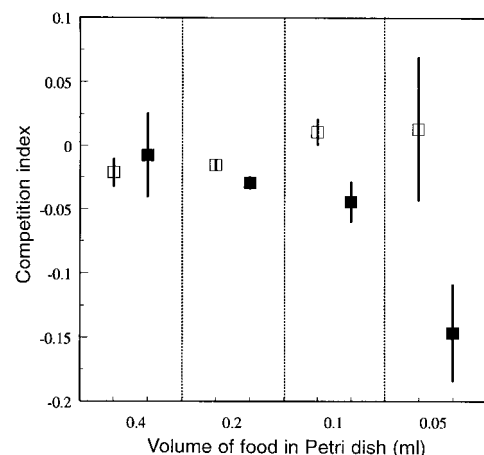


Figure 2 The competitive ability of experimental flies (control lines, open symbols; selected lines, filled symbols) relative to a tester strain, showing means and standard errors of the four selected or control lines at four levels of competition for larval food. The competition index and the statistical analysis are described in the Methods section.

resistance against pathogens and parasites in animals is scarce, and comes largely from studies on populations of farm or laboratory animals with highly modified genetic backgrounds (reviewed in ref. 1). Finally, because of the enormous amount known about *Drosophila* genetics, we suggest that the interactions between *D. melanogaster* and *A. tabida* may be a valuable model system for the study of the evolution of resistance and of the genetic basis of adaptation. □

Methods

Competitive ability of selected and control lines. Competitive ability was assessed by placing 15 second-instar larvae from the experimental lines (either control or selected insects) with 15 second-instar larvae from the tester stock (sparkling poliart, an eye-colour mutant) in agar-lined Petri dishes with variable amounts of larval medium. Four food levels were used: 0.4, 0.2, 0.1 or 0.05 ml of larval medium (25 g live baker's yeast per 100 ml water), which represent weak to strong competition regimes. We recorded the number of experimental and tester flies that survived, the development time to pupation and to adult eclosion, and female size and fluctuating asymmetry (wing length). We performed 15 replicates of each combination of line and food level.

Data analysis. Survival data were analysed using both a robust competition index (I) and a more sophisticated analysis using generalized linear modelling techniques (II). For analysis (I) we calculated the competition index¹⁵, $\log(e/(t+1))$, where e is the number of experimental and t is the number of tester flies that survived in each replicate. The means of the competition index for the eight lines were calculated and differences between selected and control lines tested using the t -test with unequal variances. For analysis (II) the untransformed survival data for the experimental flies were analysed in a generalized linear model with binomial error variances, but using quasi-likelihood estimation to account for overdispersion¹⁷. The numbers of tester flies surviving was used as a covariate and significance assessed by nesting lines within control or selection treatments (giving an F -statistic with one and six degrees of freedom). Examination of the distribution of residuals supported the choice of model. The results of the two statistical analyses were in broad agreement (values in the text are from analysis (II)), with P for (I) given first: 0.4 ml, $P > 0.1$, $P > 0.1$; 0.2 ml, $P > 0.1$, $P > 0.1$; 0.1 ml, $P = 0.033$, $P = 0.011$; 0.05 ml, $P = 0.066$, $P = 0.061$. The data for size and development time were analysed using the appropriate general linear models equivalent to analysis (II).

Received 25 March; accepted 8 July 1997.

- Read, A. F. et al. in *Ecology of Infectious Diseases in Natural Populations* (eds Grenfell, B. T. & Dobson, A. P.) 450–477 (Cambridge Univ. Press, 1995).
- Partridge, L. & Fowler, K. Direct and correlated responses to selection on age at reproduction in *Drosophila melanogaster*. *Evolution* **46**, 76–91 (1992).
- Rose, M. R. Laboratory evolution of postponed senescence in *Drosophila melanogaster*. *Evolution* **38**, 1004–1010 (1984).
- Salt, G. *The Cellular Defence Reactions of Insects* (Cambridge Univ. Press, 1970).
- Gupta, A. P. in *Comprehensive Insect Physiology, Biochemistry and Pharmacology*, Vol. 3 (eds Kerkut, G. A. & Gilbert, L. I.) 401–451 (Pergamon, Oxford, 1985).
- Strand, M. R. & Pech, L. L. Immunological basis for compatibility in parasitoid-host relationships. *Annu. Rev. Entomol.* **40**, 31–56 (1995).
- Godfray, H. C. J. *Parasitoids, Behavioral and Evolutionary Ecology*, (Princeton Univ. Press, NJ, 1994).
- Carton, Y. & Bouletreau, M. Encapsulation ability of *Drosophila melanogaster*: a genetic analysis. *Dev. Comp. Immunol.* **9**, 211–219 (1985).
- Carton, Y. & Nappi, A. J. The *Drosophila* immune reaction and the parasitoid capacity to evade it: genetic and coevolutionary aspects. *Acta Oecol.* **12**, 89–104 (1991).
- Henter, H. J. & Via, S. The potential for coevolution in a host-parasitoid system. 1. Genetic variation within an aphid population in susceptibility to a parasitic wasp. *Evolution* **49**, 427–438 (1995).
- Hughes, K. & Sokolowski, M. B. Natural selection in the laboratory for a change in resistance by *Drosophila melanogaster* to the parasitoid wasp *Asobara tabida*. *J. Insect Behav.* **9**, 477–491 (1996).
- Kraaijeveld, A. R. & van der Wel, N. N. Geographic variation in reproductive success of the parasitoid *Asobara tabida* in larvae of several *Drosophila* species. *Ecol. Entomol.* **19**, 221–229 (1994).
- Kraaijeveld, A. R. & van Alphen, J. J. M. Geographical variation in encapsulation ability of *Drosophila melanogaster* larvae and evidence for parasitoid-specific components. *Evol. Ecol.* **9**, 10–17 (1995).
- Lewontin, R. C. The effects of population density and composition on viability in *Drosophila melanogaster*. *Evolution* **9**, 27–41 (1955).
- Santos, M., Fowler, K. & Partridge, L. On the use of tester stocks to predict the competitive ability of genotypes. *Heredity* **69**, 489–495 (1992).
- Atkinson, W. D. A field investigation of larval competition in domestic *Drosophila*. *J. Anim. Ecol.* **48**, 91–102 (1979).
- McCullagh, P. & Nelder, J. A. *Generalized Linear Models* 2nd edn (Chapman & Hall, London, 1989).

Acknowledgements. We thank J. van Alphen, G. Boskamp, A. Burt, D. Ebert, J. Ellers, M. Fellowes, A. James, A. Leroi, J. McCabe, A. Orr, L. Partridge, A. Read, M. Rees and J. Werren for help and advice.

Correspondence and requests for materials should be addressed to H.C.J.G. (e-mail: c.godfray@ic.ac.uk).

Responses of primary visual cortical neurons to binocular disparity without depth perception

B. G. Cumming & A. J. Parker

University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK

The identification of brain regions that are associated with the conscious perception of visual stimuli is a major goal in neuroscience¹. Here we present a test of whether the signals on neurons in cortical area V1 correspond directly to our conscious perception of binocular stereoscopic depth. Depth perception requires that image features on one retina are first matched with appropriate features on the other retina. The mechanisms that perform this matching can be examined by using random-dot stereograms², in which the left and right eyes view randomly positioned but binocularly correlated dots. We exploit the fact that anticorrelated random-dot stereograms (in which dots in one eye are matched geometrically to dots of the opposite contrast in the other eye) do not give rise to the perception of depth³ because the matching process does not find a consistent solution. Anticorrelated random-dot stereograms contain binocular features that could excite neurons that have not solved the correspondence problem. We demonstrate that disparity-selective neurons in V1 signal the disparity of anticorrelated random-dot stereograms, indicating that they do not unambiguously signal stereoscopic depth. Hence single V1 neurons cannot account for the conscious perception of stereopsis, although combining the outputs of many V1 neurons could solve the matching problem. The accompanying paper⁴ suggests an additional function for disparity signals from V1: they may be important for the rapid involuntary control of vergence eye movements (eye movements that bring the images on the two foveae into register).

When the image of an object falls on different locations on the two retinæ, this binocular disparity gives rise to a sensation of depth (stereopsis). This process requires that an image feature on one retina be matched with an appropriate feature on the other retina, even though there are inevitably similar features nearby offering potential matches. This problem is particularly apparent in random-dot stereograms (RDS), where there are many identical dots in both images. Of course it is not necessary to consider false matches in RDS on a dot-by-dot basis—spatial filtering of the image before matching can substantially reduce the number of false matches⁵. However, this filtering alone is insufficient to solve the correspondence problem and further computation is required to eliminate false matches. Consideration of the overall pattern of disparities can indicate which potential matches form a globally consistent pattern, and many computer algorithms achieve this^{5,6}.

Although single neurons in V1 have long been known to signal the disparity of a stimulus^{7,8}, their role in stereo matching is unclear. If they are directly responsible for the conscious perception of depth, they should respond only when matches within the receptive field are registered as globally correct. If binocular neurons respond equally well to false matches at appropriate disparities, then further processing (presumably outside V1) is required to account for the psychophysical ability to discard false matches.

Whether or not disparity-selective neurons respond only to correct matches is an open question. It has been demonstrated that some V1 complex cells display disparity selectivity to dynamic RDSs and it was concluded that these neurons signal “the correct binocular matches among a multitude of false matches ... (global

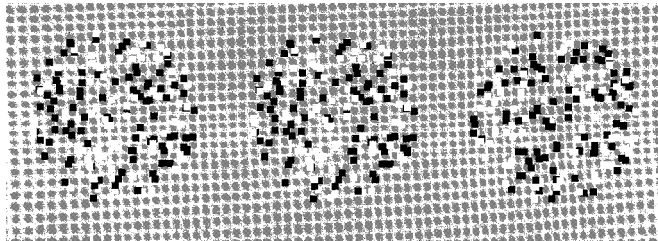


Figure 1 Example stereograms. The pair formed by the central image and the left image forms a C-RDS showing a central circular patch standing out in depth (shown for cross-eyed fusion). The pair formed by the central image and the right image is an A-RDS, and appears rivalrous with no consistent depth. Note that the right image is an exact copy of the left image in which the dot brightnesses have been reversed.

stereopsis)⁹. However, even simple local matching may yield disparity-selective responses to these stimuli, because an identical pattern of dots lies at an appropriate location on each retina when the RDS is presented at an optimal disparity. Modelling confirms this quantitatively^{10,11}, but the success of such models is not evidence against global matching in V1: all of the current experimental evidence is equally compatible with the idea that V1 neurons actually perform a more complex analysis by solving the correspondence problem.

This question requires new experimental investigations. V1 neurons should be tested with stimuli for which the predictions of local filtering models are quite different from the pattern of responses that would be expected if neurons respond exclusively to global matches. One way to do this is to present local matches in the receptive field that do not correspond to correct global matches. This can be done using anticorrelated RDS (A-RDS) because the pattern of local matches they produce appears to be rivalrous with no consistent depth², except at low dot densities (<5%)³. The perceptual process behaves as if all these local matches are false matches (matches which are discarded by global stereopsis). We therefore compared the neuronal responses to anticorrelated and correlated RDS (C-RDS) in awake behaving monkeys.

The dynamic random dot patterns were composed of 50% black and 50% white dots against a grey background (Fig. 1). An anticorrelated stereogram was obtained by inverting the contrast of the image for one eye, so that each dot shown black to the right eye was shown white to the left eye, and vice versa. This has a predictable effect on disparity selectivity for the local filtering model: contrast reversal inverts the output of each monocular subunit, so contrast reversal of only one eye's image inverts the disparity tuning (see Methods). Figure 2 shows this inversion of disparity tuning for our implementation of a model complex cell^{10,11}. So if single neurons perform only local stereoscopic matching, they should display an inverted disparity selectivity when tested with A-RDS. If they really perform global matching, they should not respond at all to A-RDS. Of course, the filter model could perhaps be elaborated so that it too could predict a failure to respond to A-RDS, but any positive experimental demonstration of disparity tuning to A-RDS must imply a failure to respond exclusively to globally correct matches.

Figure 3 shows the responses of two disparity-selective cells to dynamic A-RDS. The qualitative effects predicted by the model are

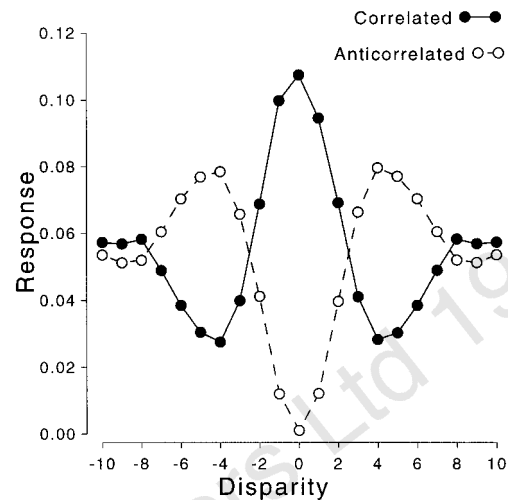


Figure 2 Responses of a model complex cell to RDS. The model neuron was tuned to zero-disparity stimuli. Filled symbols show responses to C-RDS, open symbols show responses to A-RDS. Both axes have arbitrary units. The model neuron had subunits that were matched in their spatial and temporal properties, so the two tuning curves are exact mirror images.

clearly seen: the neurons are disparity-selective to both stimuli, and the tuning curves are related by an inversion. Note that the responses to C-RDS are comparable to those reported by others^{9,12}. To quantify these effects, disparity tuning curves were fitted with Gabor functions. The relationship between the two tuning curves for a single cell was characterized by two parameters: the difference in the fitted phase, and the ratio of the fitted amplitudes. Figure 3 shows these fitted curves superimposed on the neural data, and it can be seen that they provide a good description of the main features.

Quantitative disparity tuning data for both C-RDS and A-RDS were obtained on 72 disparity-selective neurons from two animals. These data are summarized in Fig. 4, which plots the ratio of the fitted amplitudes against the difference in the fitted phases. Most results cluster around the predicted phase difference (π), but the amplitude ratio was generally smaller than the predicted value (1.0). Indeed, a number of neurons essentially did not modulate their firing with the disparity of A-RDS. By itself this does not imply that these cells have solved the correspondence problem because this result can also be explained by more complex versions of the local filtering model. Conversely, any modulation of responses to A-RDS indicates strongly that the cells respond to false local matches. As the data in Fig. 4 do not show two distinct populations of neurons, the most parsimonious interpretation of these data is that binocular V1 neurons act as some form of local filter.

The importance of this result lies in the fact that it is incompatible with the claim that single V1 neurons have "the unique capacity of solving the correspondence problem"¹³, an assertion that has not previously been critically tested. It is also incompatible with the view that these neurons perform matching dot-by-dot: in the anticorrelated stimulus, the white dots in one eye's image are uncorrelated with the white dots in the other eye's image. A scheme that matches the left edge of a white dot with the right edge of the corresponding black dot might explain the inverted responses, at least when the disparity is less than the dot width. However, such a scheme would produce responses that are not inverted for disparities greater than the dot width. We never observed such a response pattern, indeed the inverted responses were still evident at disparities much larger than the dot width (Fig. 3), as predicted by local spatial filtering.

In conclusion, most V1 neurons that show disparity selectivity for

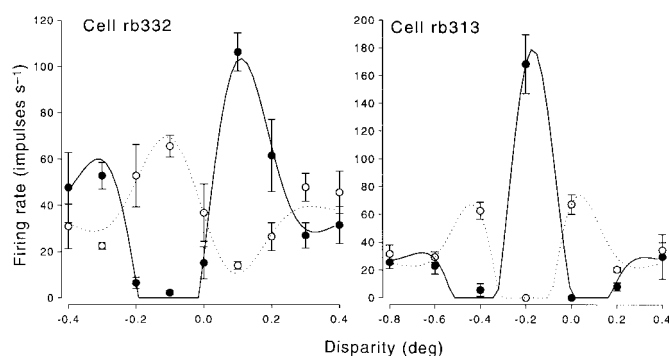


Figure 3 Responses of two complex neurons to disparity in correlated (filled symbols) and anticorrelated (open symbols) dynamic RDS. Fitted curves for the two stimulus types are Gabor functions (see Methods), in which the only parameters that are permitted to differ are the phase and the amplitude. For the cell on the left, the phase difference was 0.95π and the ratio of the amplitudes was 0.36. For the cell on the right, these values were 0.86π and 0.68, respectively. Note that although the disparity tuning curves for the two cells are quite different, the relationship between the anticorrelated responses and the correlated responses is very similar, resembling the behaviour of the model. In both cases, the tuning curve to correlated patterns is inverted by anticorrelation. The dot size was 0.08 degrees.

C-RDS also modulate their firing rate in response to disparity changes in A-RDS. The indicates that these neurons fail to distinguish reliably between global matches and false matches (local matches that are not global matches). One monkey (from whom the majority of neurons were recorded) was trained to perform psychophysical depth discrimination with RDS, but was not able to perform the discrimination with A-RDS. The fact that V1 neurons systematically alter their firing rate in response to stimulus manipulations that are not perceived psychophysically argues against a direct role for V1 neurons in binocular depth perception. As frequently suggested^{14,15}, these neurons may be involved in guiding vergence eye movements, a hypothesis that is tested in the accompanying paper examining short-latency vergence responses to A-RDS⁴. The failure to distinguish correct matches does not imply that V1 neurons have no role in stereo matching—the correct disparity matches could be extracted by combining the outputs of a number of these neurons. Our results indicate that this combination must occur outside V1. Applying the techniques described here in extrastriate cortex offers the opportunity of locating sites in the brain where stereo correspondence is achieved. □

Methods

Modelling. The model used to predict the responses of disparity-selective complex cells to dynamic RDS was our own implementation of the model described in ref. 10. Each eye's image is convolved with Gabor subunits in quadrature pairs. For each subunit, inputs from the two eyes are summed, and the binocular sum for each subunit is then squared (rectified) and summed to generate the complex cell output. The only difference between our model and that of ref. 10 was that we used positional displacements of the receptive fields to generate selectivity for different disparities, not phase differences. Both methods give similar results for RDS¹¹, and for the case of neurons tuned to zero disparity (Fig. 2) the two models are identical. For a model with N binocular subunits the response, R_c , to a correlated stimulus is given by:

$$R_c = \sum_{i=1}^N (l_i + r_i)^2 = \sum_{i=1}^N (l_i^2 + r_i^2) + 2 \sum_{i=1}^N (r_i l_i)$$

where l_i and r_i are the outputs of left and right subunits, respectively. Inverting

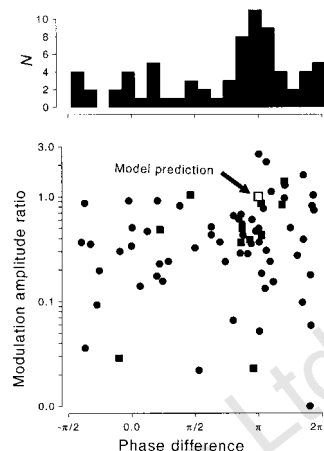


Figure 4 Summary of effect of stimulus anticorrelation on the disparity selectivity of 72 cells. Gabor functions were fitted to anticorrelated and correlated data for each cell (Fig. 3, and see Methods). In the lower panel the data from each cell are represented by a single point, which plots the relative amplitude of the fitted Gabors against the phase difference. Circles (60 cells) show results from the first animal, squares (12 cells) show results from a second. The model predicts a phase shift of π and an amplitude ratio of 1.0. Although the amplitude ratios are generally less than one (mean ratio, 0.52 ± 0.46 s.d.), most cells cluster around a phase shift of π . The upper panel shows a frequency histogram for the phase difference, confirming the existence of a cluster around π .

the contrast of (for example) the right image inverts all of the terms r_i to give R_a , the response to the anticorrelated pattern:

$$R_a = \sum_{i=1}^N (l_i - r_i)^2 = \sum_{i=1}^N (l_i^2 + r_i^2) - 2 \sum_{i=1}^N (r_i l_i)$$

The terms $\sum (l_i)^2$ and $\sum (r_i)^2$ describe the contents of the left and right images, and are not affected by changes in disparity. The term $\sum (r_i l_i)$, describes the relationship between the images in the two eyes, and gives rise to disparity selectivity. As this term is inverted by anticorrelation, the disparity tuning curve predicted by these models for anticorrelated RDS is an exact inversion of the tuning curve for correlated stimuli (Fig. 2). Note that this result does not depend upon there being a periodic receptive field: the same would be true if the subunits were simple gaussians, or any function where $f(-x) = -f(x)$.

Unit recording and stimulus presentation. Extracellular single-unit recordings were made in the primary visual cortex of two alert monkeys (*Macaca mulatta*). Scleral search coils were implanted¹⁶ in both eyes under general anaesthesia, together with a head holder and a recording chamber. Animals were then trained to maintain binocular fixation. Tungsten-in-glass micro-electrodes were introduced transdurally each day into the primary visual cortex. All procedures complied with the UK Home Office regulations on animal experimentation.

Receptive field eccentricity ranged from 1 to 4 degrees. Spike waveforms and eye-position traces were recorded to disk using the Datawave Discovery package, so that unit isolation and binocular fixation could be checked off-line. Mean firing rate was used to assess the unit response, summing all of the spikes that occurred from a time 50 ms after the first video frame of the stimulus was presented until 50 ms after the last video frame was presented.

Random-dot stimuli, like those shown in Fig. 1 (but dynamic, with a new dot pattern every 72 Hz video frame), were generated on a Silicon Graphics Indigo workstation, and displayed on two Tektronix GMA201 greyscale monitors through a haploscope. The dot width was usually 0.08 degrees (in a few simple cells it was necessary to use larger dots), the dot density was 25%, and the stimulus duration was 2 s. The stereogram always consisted of a background region at zero disparity, and a foreground region whose disparity was varied from trial to trial. Correlated and anticorrelated stimuli were alternated, but the sequence of disparities was randomized. Before testing units with RDS, the minimum response field was determined using flashing

black-and-white bars at the neuron's preferred orientation. The central region of the stereogram completely covered the minimum response field. After testing with RDS, neurons were classified as simple or complex on the basis of the modulation in their firing to drifting gratings¹⁷. Of the 72 neurons, 57 were tested in this way, of which 50 were complex cells and 7 were simple cells.

Analysis. For each neuron, the mean firing rate as a function of disparity ($f(d)$) was fitted with a Gabor function:

$$f(d) = A \exp(-(d-D)^2/2\sigma^2) \cos(2\pi\omega(d-D) + \phi) + B$$

by nonlinear regression, where A , ω and ϕ are the amplitude, spatial frequency and phase, respectively, of the cosine component, σ is the standard deviation of the gaussian, D is a position offset, and B is the baseline firing rate. In our model, this baseline firing corresponds to the activity produced by uncorrelated random-dot patterns. The correlated and anticorrelated data were fitted simultaneously, using the same values of B , ω , σ and D , but different values of A and ϕ (A_c , A_a , ϕ_c , ϕ_a , where the subscripts c and a refer to correlated and anticorrelated fits, respectively). The way in which changing from correlated to anticorrelated stimuli altered the disparity tuning of a single neurons could therefore be summarized by two parameters: an amplitude ratio A_a/A_c and a phase difference $\phi_c - \phi_a$. For the model complex cell, the amplitude ratio was 1.0 and the phase difference was π .

Received 10 February; accepted 23 June 1997.

1. Crick, F. & Koch, C. Are we aware of neural activity in primary visual cortex? *Nature* **375**, 121–123 (1995).
2. Julesz, B. *Foundations of Cyclopean Perception* (University of Chicago Press, 1971).
3. Cogan, A. I., Lomakin, A. J. & Rossi, A. Depth in anticorrelated stereograms. *Vision Res.* **33**, 1959–1975 (1993).
4. Masson, G. S., Busetini, C. & Miles, F. A. Vergence eye movements in response to binocular disparity without depth perception. *Nature* **389**, 283–286 (1997).
5. Marr, D. & Poggio, T. A computational theory of human stereo vision. *Proc. R. Soc. Lond. B* **204**, 301–328 (1979).
6. Grimson, W. E. L. A computer implementation of a theory of human stereo vision. *Phil. Trans. R. Soc. Lond. B* **292**, 217–253 (1981).
7. Barlow, H. B., Blakemore, C. & Pettigrew, J. D. The neural mechanisms of binocular depth discrimination. *J. Physiol. (Lond.)* **193**, 327–342 (1967).
8. Nikara, T., Bishop, P. O. & Pettigrew, J. D. Analysis of retinal correspondence by studying receptive fields of binocular single units in cat striate cortex. *Exp. Brain Res.* **6**, 353–372 (1968).
9. Poggio, G., Motter, B. C., Squatrito, S. & Trotter, Y. Responses of neurons in visual cortex (V1 and V2) of the alert macaque to dynamic random dot stereograms. *Vision Res.* **25**, 397–405 (1985).
10. Ohzawa, I., DeAngelis, G. C. & Freeman, R. D. Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science* **249**, 1037–1041 (1990).
11. Qian, N. Computing stereo disparity and motion with known binocular properties. *Neural Computat.* **6**, 390–404 (1994).
12. Poggio, G. F., Gonzalez, F. & Krause, F. Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *J. Neurosci.* **8**, 4531–4550 (1988).
13. Poggio, G. & Poggio, T. The analysis of stereopsis. *Annu. Rev. Neurosci.* **7**, 379–412 (1984).
14. Poggio, G. F. & Fisher, B. Binocular interactions and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J. Neurophysiol.* **40**, 1392–1405 (1977).
15. Poggio, G. Mechanisms of stereopsis in monkey visual cortex. *Cerebr. Cort.* **3**, 193–204 (1995).
16. Judge, S. J., Richmond, B. J. & Chu, F. C. Implantation of magnetic search coils for measurement of eye position: an improved method. *Vision Res.* **30**, 535–538 (1980).
17. Skottun, B. C. *et al.* Classifying simple and complex cells on the basis of response modulation. *Vision Res.* **31**, 1079–1086 (1991).

Acknowledgements. This work was supported by the Wellcome Trust and the Royal Society.

Correspondence and requests for materials should be addressed to B.G.C. (e-mail: bc@physiol.ox.ac.uk).

Vergence eye movements in response to binocular disparity without depth perception

G. S. Masson*, C. Busetini & F. A. Miles

Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

*Centre de Recherche en Neurosciences Cognitives, Centre National de la Recherche Scientifique, 13402 Marseille, France

Primates use vergence eye movements to align their two eyes on the same object and can correct misalignments by sensing the difference in the positions of the two retinal images of the object (binocular disparity). When large random-dot patterns are viewed dichoptically and small binocular misalignments are suddenly imposed (disparity steps), corrective vergence eye

movements are elicited at ultrashort latencies^{1,2}. Here we show that the same steps applied to dense anticorrelated patterns, in which each black dot in one eye is matched to a white dot in the other eye, initiate vergence responses that are very similar, except that they are in the opposite direction. This sensitivity to the disparity of anticorrelated patterns is shared by many disparity-selective neurons in cortical area V1 (ref. 3), despite the fact that human subjects fail to perceive depth in such stimuli^{4,5}. These data indicate that the vergence eye movements initiated at ultrashort latencies result solely from locally matched binocular features, and derive their visual input from an early stage of cortical processing before the level at which depth percepts are elaborated.

Disparity-selective neurons have often been implicated in the perception of depth (stereopsis)⁶, and it has been shown that in the first stage of the cortical visual pathways in area V1, such neurons are sensitive to the disparity of anticorrelated patterns³, even though such patterns are perceptually rivalrous, cannot be fused, and lack consistent depth^{4,5}. Furthermore, the disparity tuning curves of the neurons were often inverted with anticorrelated patterns, a characteristic of simple local filtering models^{7,8}. These findings are consistent with the hypothesis that these neurons respond to purely local matches between the images seen by the two eyes, regardless of whether a global match is present⁹. Thus, such neurons do not solve the correspondence problem and can represent only a rudimentary stage in the processing of binocular signals for stereopsis. We now provide evidence that such rudimentary binocular signals can generate motor responses by showing that small disparity stimuli applied to dense anticorrelated patterns give rise to inverted vergence eye movements at ultrashort latencies.

A variety of cues can be used to control the angle of convergence between the two lines of sight^{10,11}, but the only cue of concern here is binocular disparity^{1,12}, which provides a direct measure of the misalignment of the two eyes with respect to the object(s) of interest (vergence error) and is assumed to be sensed directly by disparity-selective neurons^{13,14}. Examples of the initial vergence responses elicited by small horizontal disparities ($<2^\circ$) applied to large correlated random-dot patterns (matching images at the two eyes) are seen in Fig. 1 (continuous line), which shows mean vergence velocity profiles for one human (Fig. 1a) and for one monkey (Fig. 1c). Stimuli were presented on a tangent screen using two slide projectors and orthogonal polarizing filters to allow independent control of the images seen by each eye. Each trial started with the screen blank and then stationary patterns with a given horizontal disparity were presented. For the data shown in Fig. 1a, c, all patterns had crossed disparities (the pattern seen by the right eye had been shifted leftwards; the pattern seen by the left eye had been shifted rightwards), simulating the abrupt appearance of a textured surface in front of the tangent screen. Such stimuli initiated increased convergence—the correct response to restore binocular alignment—with a latency of <60 ms in the case of the monkey and <90 ms in the case of the human^{1,2}. The initial vergence responses to anticorrelated random-dot patterns with similar crossed disparities are shown as dotted lines in Fig. 1a, c and are clearly in the reverse direction. These inverted responses have a comparably short latency but a slightly lower rate of acceleration.

We quantified these initial motor responses by measuring the change in vergence position over a 33-ms period commencing at a fixed time after the appearance of the disparity stimuli: 60 ms for the monkey, 90 ms for the human. This meant that our measures were restricted to the initial (open-loop) vergence responses that were generated by the disparity input before it had been affected by eye-movement feedback. Disparity tuning curves based on these measures are plotted in Fig. 1b (human) and Fig. 1d (monkey), and show the characteristic S-shapes with non-zero asymptotes¹, consistent with the operation of a depth-tracking servo of modest range. Thus, with normal (correlated) patterns and disparities up to a degree or two, the slopes are positive and small increases in the

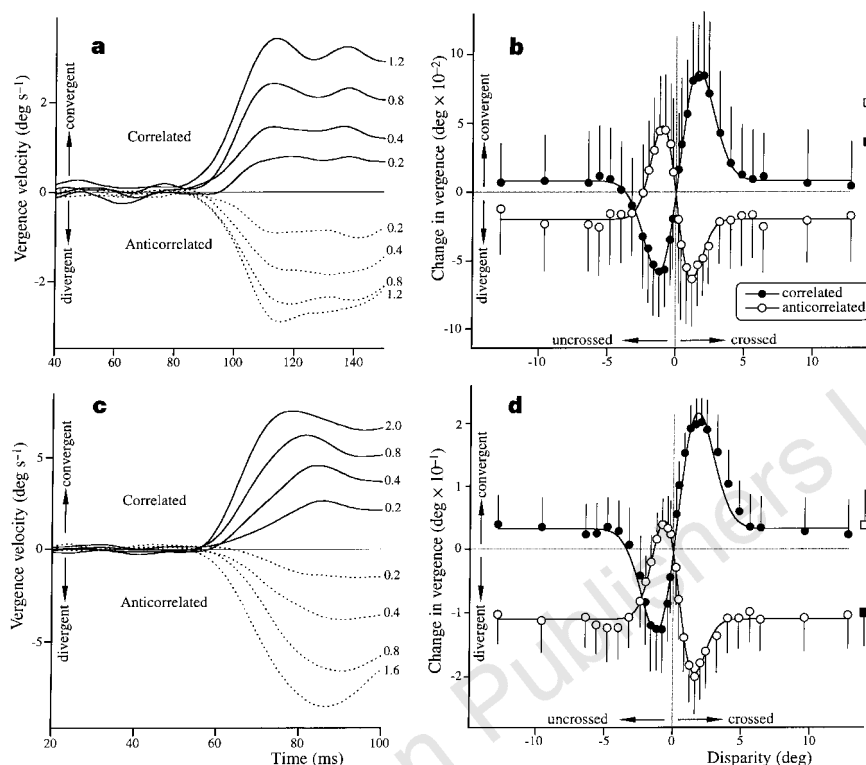


Figure 1 Vergence eye movements elicited by disparity stimuli applied to high-density random-dot patterns (50%). **a**, Mean vergence velocity responses of human subject, F.A.M., to crossed disparity stimuli applied to correlated (continuous line) and anticorrelated (dotted line) patterns, with stimulus magnitudes (in degrees) at the ends of traces. Note, vergence velocity is the difference between velocity of the two eyes (left eye–right eye) and increasing convergence is positive (upward deflection). **b**, Plot of mean ($\pm$ s.d.) changes in vergence position (over period 90–123 ms, starting from stimulus onset) against

disparity stimulus for human subject, F.A.M., with correlated (filled circles) and anticorrelated (open circles) patterns; also shown are control responses to zero-disparity stimuli applied to correlated (filled square) and anticorrelated (open square) patterns. **c**, Mean vergence velocity responses of monkey, Bo, in response to crossed disparity stimuli. **d**, Plot of mean ($\pm$ s.d.) changes in vergence (over period 60–93 ms from stimulus onset) for monkey, Bo. Individual traces and datum points based on at least 178 responses.

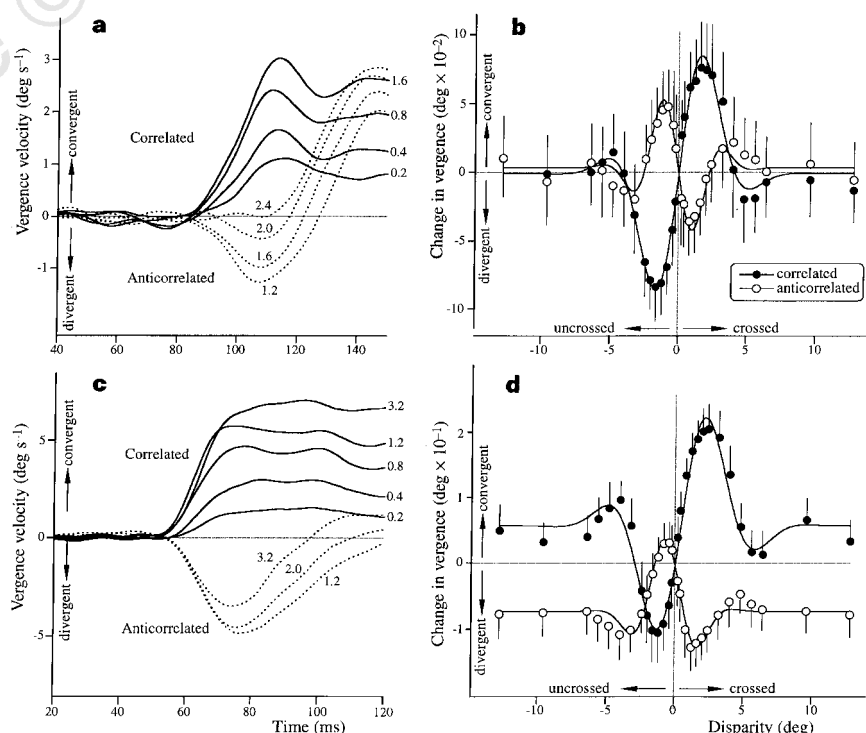


Figure 2 Vergence eye movements elicited by disparity stimuli applied to low-density random-dot patterns (7.5%). Layout, subjects and conventions all identical to Fig. 1. Individual traces and datum points based on at least 84 responses.

Table 1 Best-fit Gabor parameters

Subject	<i>A</i>	σ	<i>D</i>	ω	ϕ	<i>B</i>	<i>R</i>
Correlated patterns							
F.A.M.	0.39	1.48	0.27	0.0337	271.6	0.008	0.146
R.J.K.	0.35	1.69	0.20	0.0353	272.3	0.002	0.156
Bo	0.93	1.53	0.32	0.0324	271.2	0.032	0.346
Lu	0.94	1.58	0.12	0.0189	270.4	-0.002	0.212
Ch	0.49	1.84	1.33	0.0667	278.8	0.150	0.420
Anticorrelated patterns							
F.A.M.	0.58	1.08	-0.02	0.0234	88.0	-0.020	0.112
R.J.K.	0.45	1.13	0.30	0.0198	89.1	-0.024	0.077
Bo	0.83	1.12	0.29	0.0336	85.9	-0.111	0.241
Lu	0.53	1.09	-0.25	0.0270	85.3	-0.014	0.123
Ch	1.23	1.03	0.95	0.0250	86.3	-0.153	0.249

Best-fit parameters when the following Gabor function was fitted to the disparity tuning curves obtained with high-density random-dot patterns (50%):

$$f(d) = A \exp\left(-\frac{(d-D)^2}{2\sigma^2}\right) \cos(2\pi\omega(d-D) + \phi) + B$$

where *d* is the stimulus disparity, *A* is a gain factor, σ is the gaussian width, ω and ϕ are the spatial frequency and phase of the cosine term, *D* is the displacement, and *B* is an offset parameter to allow for the non-zero asymptotes. An iterative procedure with a least-squares criterion was used to obtain the best fits with each of the parameters resolved to the number of decimal places shown. *R* is the peak-to-peak amplitude derived from the best-fit Gabor functions. All units are in degrees except for ω , which is in cycles per deg. Data are for two human subjects (F.A.M., R.J.K.) and three monkeys (Bo, Lu, Ch).

magnitude of the stimulus (disparity) lead to roughly proportional increases in the magnitude of the response (vergence) in the appropriate direction: crossed disparities elicit increased convergence and uncrossed disparities decreased convergence. Larger disparities exceed the system's operating range and responses default towards a non-zero level that is idiosyncratic and might reflect residual correlations¹. In contrast, the disparity tuning curves obtained with anticorrelated patterns are almost mirror images of those obtained with correlated patterns. Similar clear evidence for a sign inversion with anticorrelated patterns was obtained for two more monkeys and one more human.

The curves in Fig. 1b,d are least-squares-fit Gabor functions, with an offset term to allow for the non-zero asymptotes. Such functions are commonly used in vision research to characterize linear filters and they provide a good fit to our data, allowing an objective quantitative comparison of the responses to correlated and anticorrelated patterns. The parameters for these fits, together with those for the additional subjects, are listed in Table 1. To give some estimate of the amplitude of modulation with disparity, Table 1 includes the peak-to-peak amplitude of the Gabor functions (*R*). (This derived measure was used because it describes the amplitude of modulation reliably, independent of any one term in the fitted function. Note that the spatial frequencies, ω , are always very low so that, within the important disparity range, $\pm 5^\circ$, the cosine terms are almost linear and pass very close to the origin, with positive slopes for the correlated data and negative slopes for the anticorrelated data.) The amplitude measure (*R*) and the gaussian width (σ) were invariably smaller with the anticorrelated patterns than they were with the correlated patterns, on average by 37% (range, 23–51%) and 32% (range, 27–44%), respectively. The sign inversion is evident from the difference in the phase of the cosine terms ($\phi_{\text{correlated}} - \phi_{\text{anticorrelated}}$), which generally approximated π : 183.6° (human subject F.A.M.), 183.2° (human R.J.K.), 185.3° (monkey Bo), 185.1° (monkey Lu) and 192.5° (monkey Ch). The data for the two species were similar in all essentials, indicating that the monkey is a good animal model for the human.

The preceding paper³ shows that many neurons in V1 respond to the local matches in anticorrelated patterns, often with a sign inversion, and we have shown here that these patterns can elicit vergence eye movements with a sign inversion at ultrashort latencies. We think it likely that the two observations are causally linked: that is, the earliest vergence eye movements derive their visual input

from an early stage of cortical processing, possibly even as early as V1 (ref. 15). Cortical disparity-selective neurons that might drive vergence have been classified as 'near', 'far', 'tuned-near' and 'tuned-far', depending on the range of disparities over which they are active¹⁴. The vergence response to a given disparity in our experiments is determined by the net balance of activity in the population of these disparity-selective neurons that influences the vergence state. The preceding paper³ helps to explain how this balance is likely to differ with correlated and anticorrelated patterns. For example, consider the neurons like those in Fig. 3 of ref. 3, whose disparity tuning curves are inverted with anticorrelated patterns (and whose Gabor fits show a phase difference of roughly π). The subset of such neurons that is maximally active when a given disparity is applied to the correlated patterns, for example, will be minimally active when that same disparity is applied to the anticorrelated patterns (and vice versa). In fact, the balance of activity within the whole population of such neurons when a given disparity is applied to an anticorrelated pattern will be the mirror image of that obtained when the same disparity is applied to a correlated pattern. The net result is to change the sign of the disparity feedback signal driving vergence: if the balance of activity when a given disparity is applied to the correlated pattern results in increased convergence, then the balance when that same disparity is applied to the anticorrelated pattern would result in decreased convergence and vice versa. Such neurons might therefore explain our inverted vergence data with anticorrelated patterns, although we cannot rule out a contribution from neurons that show phase shifts of less than π with anticorrelated stimuli. Note that, because we are dealing with a population response, the Gabor parameters for vergence eye movements need not be expected to match the Gabor parameters for any given cell type.

From the functional standpoint, the sensitivity to anticorrelated patterns indicates that the sensory mechanisms underlying the vergence eye movements studied here deal solely with local disparity matches, regardless of the global disparity match, and consequently do not solve the correspondence problem. This is in line with our previous suggestion that these reflex-like vergence mechanisms normally function only to eliminate small (residual?) vergence errors, merely passively aligning the two eyes on the nearest available salient object(s)¹. Additional mechanisms are needed to transfer gaze actively to new depth planes, which in a crowded visual world requires target selections that must often pose a correspondence problem. This seems likely to involve higher-level processing and to require much more time, perhaps accounting for the much longer latencies—160 ms or more—generally reported for vergence eye movements^{12,16,17}. Evidence for an intermediate stage of processing comes from some additional experiments that we have done with low-density random-dot patterns. All of the data described so far were obtained with random dots that covered 50% of the image space, but we also tested four subjects with patterns in which the dots were more sparsely distributed. Reducing the dot density to 7.5% (by reducing the number of dots) had little effect on the vergence eye movements elicited with correlated patterns, but the inverted responses recorded with anticorrelated patterns were now very transient (see Fig. 2: same layout as for Fig. 1). In fact, with these anticorrelated patterns, the vergence velocity profiles were now biphasic, the initial reversal lasting only 30–60 ms and being replaced by a 'correctly' directed response. The latter ranged from robust (as in Fig. 2a) to very weak (as in Fig. 2c), and presumably resulted from a later, more global stage of processing by neurons that respond to spatial matching independently of contrast. When dot density was further reduced to 0.6% (by also reducing dot size) these later 'correctly' directed responses became a consistent feature, whereas the initial transient reversals were seen in only 3 of the 8 data sets (not shown). This is in accord with our well known ability to initiate appropriately directed vergence eye movements to small single-line stimuli that have opposite contrast at the two eyes^{16,18}.

In a two-alternative forced-choice procedure, three human observers (two of whom were naive) were able correctly to discriminate 1.2° crossed and uncrossed disparities applied to our correlated patterns, regardless of dot density. None of the subjects was able to make these discriminations with our denser anticorrelated patterns (50 and 7.5%), even when a central (binocular) fixation spot remained available on the screen as a reference to allow a relative depth judgement¹⁹. However, all of the subjects were able to make these discriminations (correctly) with the least-dense anticorrelated patterns (0.6%), with or without the reference spot. These findings are all in agreement with a previous study⁵ which showed that some subjects can perceive depth in low-density anticorrelated patterns (<5%). Our data with dense patterns provide a clear instance of a dissociation between sensory perception (depth) and motor responses (short-latency vergence). This is consistent with the idea that the earliest vergence responses reported here depend on inputs derived from an early stage of cortical processing. □

Methods

Vergence eye movements. Most techniques have been described previously^{1,20}. The positions of both eyes were recorded using the electromagnetic search coil technique²¹. Subjects faced a tangent screen (viewing distance 33 cm; subtense, 80° × 80°) onto which two random-dot patterns were simultaneously back-projected. Patterns could be (1) high-density (2° diameter dots covering 50% of the image space) or (2) low-density (2° dots with 7.5% coverage, or 0.5° dots with 0.6% coverage), with the additional constraint that dot centres had minimal separations of 5°. Orthogonal polarizing filters in the projection paths ensured that each of the two eyes saw only one of the patterns, the horizontal positioning of which was controlled by mirror galvanometers. For high-density correlated stimuli, patterns seen by each eye had matching black dots on a white background. For high-density anticorrelated stimuli, the left eye saw black dots on a white background and the right eye saw a matching negative image (white dots on a black background). The low-density patterns were similarly arranged, except that the dots always appeared against a grey background. Trials started with the screen blank (same space-averaged luminance as for the patterns), except for a target spot projected onto the screen 10° right of centre, which the subject was required to fixate. After a randomized interval this spot was extinguished and a second appeared at the centre of the screen. Subjects were required to make a saccadic eye movement to acquire this new target, at which time the target was switched off. With gaze now directed at the screen centre, stationary random-dot patterns with a fixed disparity appeared (post-saccadic delay: 30 ms for monkeys and 50 ms for humans) for a brief period (100 ms for monkeys and 200 ms for humans) before the screen was blanked, ending the trial. This procedure served to apply the disparity stimuli in the wake of centring saccades to take advantage of post-saccadic enhancement^{1,2}. Disparities ranged from 0 to 12.8° (crossed and uncrossed, correlated and anticorrelated patterns) and varied randomly from trial to trial. All data shown have had the responses to zero disparities (plotted separately as square symbols in Fig. 1b, c) subtracted to eliminate any post-saccadic vergence drifts and idiosyncratic responses to the mere appearance of a pattern. This has the effect of forcing all the disparity tuning curves through the origin.

Psychophysical tests of depth discrimination. In a two-alternative forced-choice procedure, human subjects were asked to indicate whether a random-dot pattern appeared nearer or farther from the projection screen when subjected to 1.2° crossed and uncrossed disparities for 200 ms, exactly as in the experiments described above. Because our patterns provide only absolute disparity cues, trials were also included in which a central fixation spot remained available as a reference¹⁹.

Received 10 February; accepted 27 June 1997.

1. Busetini, C., Miles, F. A. & Krauzlis, R. J. Short-latency disparity vergence responses and their dependence on a prior saccadic eye movement. *J. Neurophysiol.* **75**, 1392–1410 (1996).
2. Busetini, C., Miles, F. A. & Krauzlis, R. J. Short-latency disparity vergence responses in humans. *Soc. Neurosci. Abstr.* **20**, 1403 (1994).
3. Cumming, B. G. & Parker, A. J. Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature* **389**, 280–283 (1997).

4. Julesz, B. Binocular depth perception of computer-generated patterns. *Bell System Tech. J.* **39**, 1125–1162 (1960).
5. Cogan, A. I., Lomakin, A. J. & Rossi, A. F. Depth in anticorrelated stereograms: Effects of spatial density and interocular delay. *Vision Res.* **33**, 1959–1975 (1993).
6. Bishop, P. O. & Pettigrew, J. D. Neural mechanisms of binocular vision. *Vision Res.* **26**, 1587–1600 (1986).
7. Ohzawa, I., DeAngelis, G. C. & Freeman, R. D. Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science* **249**, 1037–1041 (1990).
8. Qian, N. Computing stereo disparity and motion with known binocular properties. *Neural Comput.* **6**, 390–404 (1994).
9. Parker, A. J. & Cumming, B. G. Local vs global stereoscopic matching in neurons of cortical area V1. *Invest. Ophthalmol. Vis. Sci.* **37**, S424 (1996).
10. Collewijn, H. & Erkelens, C. J. in *Eye Movements and their Role in Visual and Cognitive Processes* (ed. Kowler, E.) 213–261 (Elsevier, Amsterdam, 1990).
11. Judge, S. J. in *Vision and Visual Dysfunction, 8: Eye Movements* (ed. Carpenter, R. H. S.) 157–172 (Macmillan, London, 1991).
12. Rashbass, C. & Westheimer, G. Disjunctive eye movements. *J. Physiol. (Lond.)* **159**, 339–360 (1961).
13. Poggio, G. F. & Fischer, B. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J. Neurophysiol.* **40**, 1392–1405 (1977).
14. Poggio, G. F. Mechanisms of stereopsis in monkey visual cortex. *Cerebr. Cort.* **5**, 193–204 (1995).
15. Trotter, Y., Celebrini, S., Stricanne, B., Thorpe, S. & Imbert, M. Neural processing of stereopsis as a function of viewing distance in primate visual cortical area V1. *J. Neurophysiol.* **76**, 2872–2885 (1996).
16. Mitchell, D. E. Properties of stimuli eliciting vergence eye movements and stereopsis. *Vision Res.* **10**, 145–162 (1970).
17. Cumming, B. G. & Judge, S. J. Disparity-induced and blur-induced convergence eye movement and accommodation in the monkey. *J. Neurophysiol.* **55**, 896–914 (1986).
18. Westheimer, G. & Mitchell, D. E. The sensory stimulus for disjunctive eye movements. *Vision Res.* **9**, 749–755 (1969).
19. Regan, D., Erkelens, C. J. & Collewijn, H. Necessary conditions for the perception of motion in depth. *Invest. Ophthalmol. Vis. Sci.* **27**, 584–597 (1986).
20. Busetini, C., Miles, F. A., Schwarz, U. & Carl, J. R. Human ocular responses to translation of the observer and of the scene: dependence on viewing distance. *Exp. Brain Res.* **100**, 484–494 (1994).
21. Fuchs, A. F. & Robinson, D. A. A method for measuring horizontal and vertical eye movement chronically in the monkey. *J. Appl. Physiol.* **21**, 1068–1070 (1966).

Acknowledgements. G.S.M. was supported by La Fondation pour la Recherche Medicale (France).

Correspondence and requests for materials should be addressed to F.A.M. (e-mail: fam@lsr.nei.nih.gov).

Action-potential propagation gated by an axonal I_A-like K⁺ conductance in hippocampus

Dominique Debanne*, Nathalie C. Guérineau*, Beat H. Gähwiler & Scott M. Thompson

Brain Research Institute, August Forel-Strasse 1, CH-8029 Zurich, Switzerland

Integration of membrane-potential changes is traditionally reserved for neuronal somatodendritic compartments. Axons are typically considered to transmit reliably the result of this integration, the action potential¹, to nerve terminals^{2,3}. By recording from pairs of pyramidal cells in hippocampal slice cultures^{4–6}, we show here that the propagation of action potentials to nerve terminals is impaired if presynaptic action potentials are preceded by brief or tonic hyperpolarization. Action-potential propagation fails only when the presynaptic action potential is triggered within the first 15–20 ms of a depolarizing step from hyperpolarized potentials; action-potential propagation failures are blocked when presynaptic cells are impaled with electrodes containing 4-aminopyridine, indicating that a fast-inactivating, A-type K⁺ conductance is involved. Propagation failed between some, but not all, of the postsynaptic cells contacted by a single presynaptic cell, suggesting that the presynaptic action potentials failed at axonal branch points. We conclude that the physiological activation of an I_A-like potassium conductance can locally block propagation of presynaptic action potentials in axons of the central nervous system. Thus axons do not always behave as simple electrical cables: their capacity to transmit action potentials is determined by a time-dependent integration of recent membrane-potential changes.

* Present addresses: Unité de Neurocybernétique Cellulaire, UPR9041 CNRS, 280 Bld Sainte Marguerite, 13009 Marseille, France (D.D.); CCIPE, U469 INSERM, 141 rue de la Cardonille, 34094 Montpellier CEDEX 5, France (N.C.G.).

Intracellular recordings were made from pairs of monosynaptically coupled hippocampal pyramidal cells^{4,5} in organotypic slice cultures⁶. The postsynaptic cell can be used as a detector for the propagation of action potentials in the presynaptic cell axon, provided there are no complete failures of transmitter release from all of the synapses formed by that axon with the postsynaptic cell. Only monosynaptic connections⁵, which have relatively invariant latencies between the peak of the presynaptic action potential (pAP) and the onset of the excitatory postsynaptic potential or current (EPSP/C), were studied (Fig. 1a, top right). Cell pairs were studied only if EPSP/Cs were elicited in every trial (>50) when the pAP was evoked with a brief depolarization at 0.3 Hz from membrane potentials less negative than -65 mV (Fig. 1a, lower right histogram). If a brief hyperpolarizing current pulse was applied immediately before induction of the pAP, no EPSP was observed (Fig. 1b). When the interval between the end of the hyperpolarizing current pulse and the induction of the pAP was >20 ms, then the pulse did not prevent the pAP from eliciting an EPSP (Fig. 1c). Similar effects were observed in four of nine CA3–CA3 and four of thirteen CA3–CA1 cell pairs tested, and persisted for as long as the two impalements could be maintained (up to 2.5 h; mean duration, 56 ± 21 min). Small, locally elicited hyperpolarizing inhibitory synaptic potentials were also sufficient to prevent subsequently evoked pAPs from eliciting EPSPs in many trials ($n = 3$ cell pairs; results not shown).

Failure to observe an EPSP/C after a hyperpolarizing current pulse can be attributed to failure of pAP propagation to the nerve terminal, rather than failure of the pAP to trigger transmitter release. First, there were no failures when the pAP was evoked at potentials less negative than -65 mV. Second, EPSPs failed in an all-or-none fashion (compare with ref. 7). The distribution of non-failure EPSP amplitudes was similar for interstimulus intervals that resulted in intermittent failures and for EPSPs elicited with no hyperpolarizing pulse (Fig. 1d). A direct effect of the hyperpolarizing pulse on the membrane potential of the presynaptic terminal cannot be ruled out, but it is electrotonically unlikely because CA3 cell axons are unmyelinated⁸, the hyperpolarizing pulses were small (15 – 40 mV), and the distance between pre- and postsynaptic cell somata was large (>1 mm for CA3–CA1 and >150 μ m for CA3–CA3 connections). Furthermore, synaptic responses could not be elicited in the presence of tetrodotoxin (compare with ref. 7).

We considered that propagation failures occur when the depolarizing pulse used to elicit the pAP activates a voltage-dependent conductance that is inactivated at the resting membrane potential (RMP). Because the hyperpolarizing pulse does not block the action potential recorded with the somatic electrode in the presynaptic cell, this conductance must shunt the action potential in the axon at some point distal to its site of initiation in or near the axon hillock^{9,10}.

Propagation failures were also observed when the pAP was elicited from a constant hyperpolarized potential (Fig. 2). The probability of observing propagation failures depended upon the interval between the onset of the depolarizing pulse and the occurrence of the pAP. Propagation failures were not observed with delays of $>20 \pm 4$ ms ($n = 6$) (Fig. 2a), implicating a conductance with fast inactivation. Recovery from inactivation occurred rapidly at potentials more negative than -80 mV, as tested with a subthreshold depolarizing prepulse (30 ms duration) elicited at various intervals preceding the pAP. When the prepulse preceded the pAP by $>22 \pm 7$ ms ($n = 3$), propagation failures did not occur (Fig. 2b).

The involvement of I_A in propagation failure was tested using microelectrodes containing 4-aminopyridine (40 mM), a blocker of I_A -like conductances^{11,12}, to record from the presynaptic cell. Upon impalement, pAPs failed to elicit EPSPs when the cells were hyperpolarized. Within 2 min, however, the pAP always elicited an EPSP ($n = 5$) (Fig. 3a). 4-Aminopyridine had no effect on the

amplitude of EPSPs elicited from the RMP ($103 \pm 11\%$; $n = 3$). We conclude that activation of an I_A -like conductance underlies propagation failure. Furthermore, the rapidity with which 4-aminopyridine blocks propagation failure suggests that this conductance is located within a short distance of the cell body. I_D is also sensitive to 4-aminopyridine¹³, but its inactivation and recovery from inactivation are too slow to be involved in the effects described.

Because inactivation of I_A is fast, bursts of pAPs generated at hyperpolarized potentials should be transmitted more reliably than single pAPs. Indeed, although the first pAP in a train often failed to trigger an EPSP, subsequent pAPs regularly elicited EPSPs ($n = 4$) (Fig. 3b).

Simulations of action-potential propagation in axons predict that conduction can be delayed by ~ 1 ms or abolished when membrane conductance is locally increased¹⁴. We therefore measured the synaptic latency, which is the sum of the axonal pAP conduction time and the presumably invariant synaptic delay. Indeed, synaptic latency was greatest for pAPs elicited with the briefest delays that did not result in propagation failures (Fig. 4a). On average, synaptic latency increased by 0.8 ± 0.1 ms ($n = 5$). These results are thus

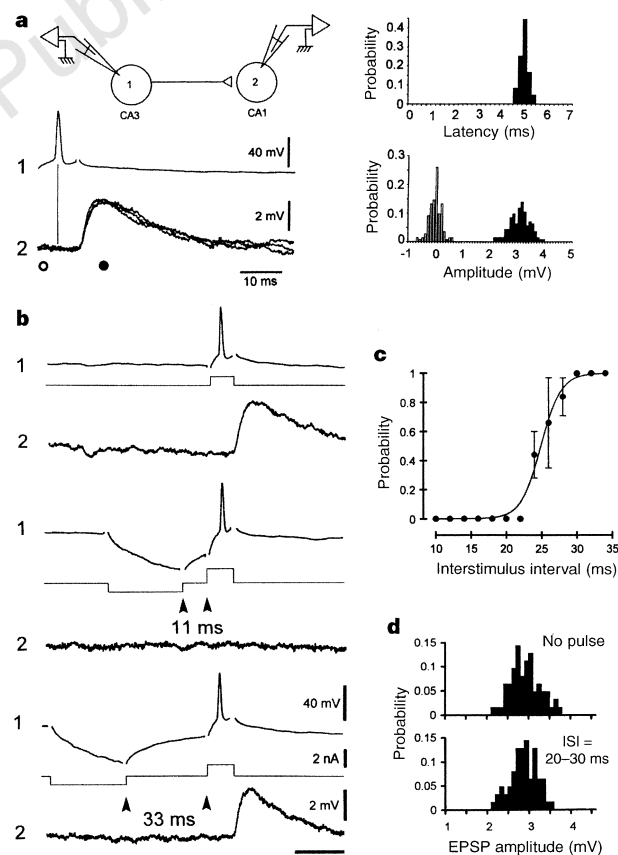


Figure 1 Axon-potential propagation failures. **a**, With cell (1) at -65 mV, presynaptic action potentials (pAP) evoked EPSPs in cell (2) at fixed latency (upper right histogram), without failures (lower right histogram, $n = 142$ trials). EPSP amplitudes (right-hand set of bars) and recording noise (left-hand set of bars) measured at fixed latencies (indicated by $\circ$ and $\bullet$). **b**, In the same pair of cells, EPSPs were elicited only when a hyperpolarizing pulse preceded the pAP by a long interval. The interval is indicated by arrowheads. **c**, Probability of EPSP occurrence as a function of interval between end of pre-pulse and pAP (sigmoid function: $X_{50} = 25$ ms; $r^2 > 0.99$; $n = 250$ trials). **d**, Same pair of cells: histograms of non-failure EPSP amplitudes in the absence of hyperpolarizing pre-pulses (no pulse) and at the threshold for propagation failures (ISI, interstimulus interval, 20–30 ms).

consistent with the hypothesis that depolarization of the presynaptic cell from a hyperpolarized potential activates a shunting conductance in the cell's axon.

Failures of propagation were not observed in all cell pairs. This could indicate that the I_A -like conductance is not sufficiently large to shunt the pAP in every cell, or it may reflect anatomical and geometrical differences between different axon collaterals of a single presynaptic cell. In three cell pairs in which hyperpolarization of the presynaptic cell prevented the occurrence of EPSPs in one postsynaptic cell, another postsynaptic cell could be impaled in which propagation failures could not be generated (Fig. 4b). The pAP can thus fail to propagate in some, but not all, axon collaterals upon activation of I_A .

We conclude that there are I_A -like K^+ channels in the proximal axonal membrane that are inactivated at somatic membrane potentials less negative than -65 mV, permitting action potentials to propagate reliably to nerve terminals. When the presynaptic neuron is transiently hyperpolarized, such as immediately after an inhibitory synaptic potential, the I_A channels are deinactivated. They can then become rapidly activated during a depolarization preceding the occurrence of an action potential, and shunt propagation of the action potential along the axon.

Axonal branch points and swellings are associated with failures of action-potential propagation^{15,16}. The axons of CA3 pyramidal cells often have many branch points close to the cell body^{8,17}, and we suggest that pAP propagation failures may occur there. There is immunocytochemical evidence that the A-type K^+ channel Kv1.4 is

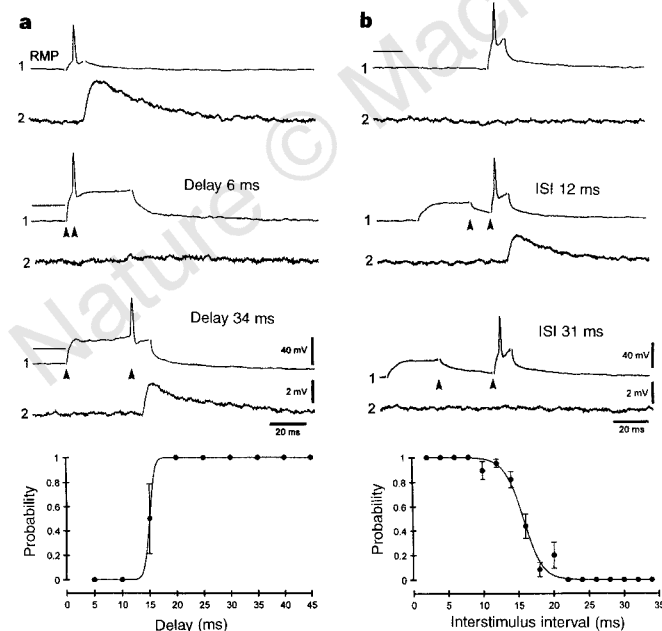


Figure 2 Inactivation and deinactivation of conductance underlying failure of action potential (AP) propagation. Same cell pair as in Fig. 1. **a**, At the resting membrane potential (RMP) (-65 mV; horizontal line), pAPs in cell (1) always evoked EPSPs in cell (2). With cell (1) at -100 mV, EPSPs occurred only when pAPs were elicited with long delays, as summarized in the bottom panel (80 trials, sigmoid function: $X_{50} = 15$ ms; $r^2 > 0.999$). Delay time is indicated by arrowheads. **b**, Presynaptic cell at -85 mV. Propagation failures occurred only when a depolarizing pulse preceded AP induction by a long interval, as summarized in the bottom panel (80 trials, sigmoid function, $X_{50} = 16$ ms; $r^2 > 0.98$). ISI, interstimulus interval, indicated by arrowheads.

present in the axons of hippocampal pyramidal cells¹⁸ and electrophysiological evidence for 4-aminopyridine-sensitive K^+ channels in the axons of retinal ganglion cells¹⁹.

The frequency-filtering characteristics of the hippocampus will be significantly affected by the phenomena described here. If the membrane potential of pyramidal cells is more negative than -65 mV, for example as a result of hyperpolarizing GABAergic synaptic potentials or the action of extrinsic neuromodulators, then I_A is available for activation by all fast depolarizations, such as EPSPs. Suprathreshold EPSPs will trigger action potentials, but their propagation will be impaired because the time-to-peak of EPSPs (5–10 ms) is shorter than that required for the channels to inactivate (~ 24 ms). Tonic hyperpolarization thus favours transmission of bursts of action potentials (Fig. 3b). Cell-to-cell transmission of single action potentials, in contrast, will occur best when pyramidal cells are relatively depolarized.

The temporal sequence of membrane potential changes preced-

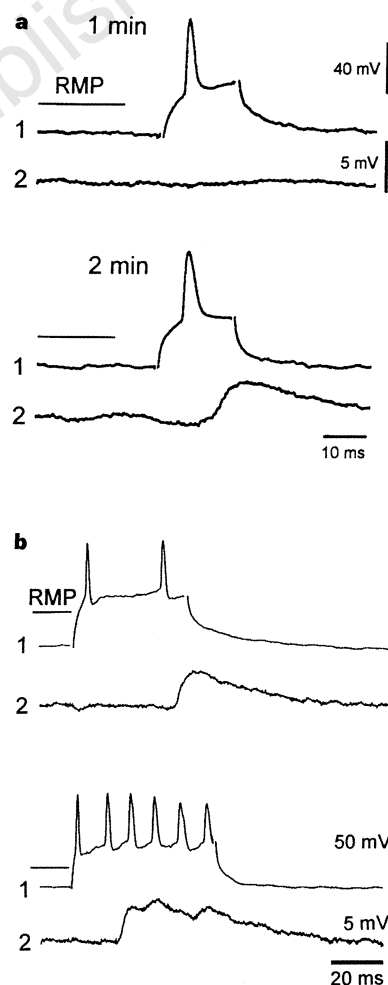


Figure 3 I_A mediates propagation failure. **a**, When presynaptic cells (1) were impaled with electrodes containing 40 mM 4-aminopyridine, the voltage-dependent failure to elicit EPSPs was observed during the first minute of recording, but not when 4-aminopyridine had diffused into the presynaptic cell for 2 min. The resting membrane potential (RMP) of cell 1 was -62 mV. **b**, When two or more action potentials were elicited from potentials more negative than -75 mV, the first action potential always failed to trigger an EPSP, but no failures were seen following the second or third action potential in a burst ($n, 5-12$).

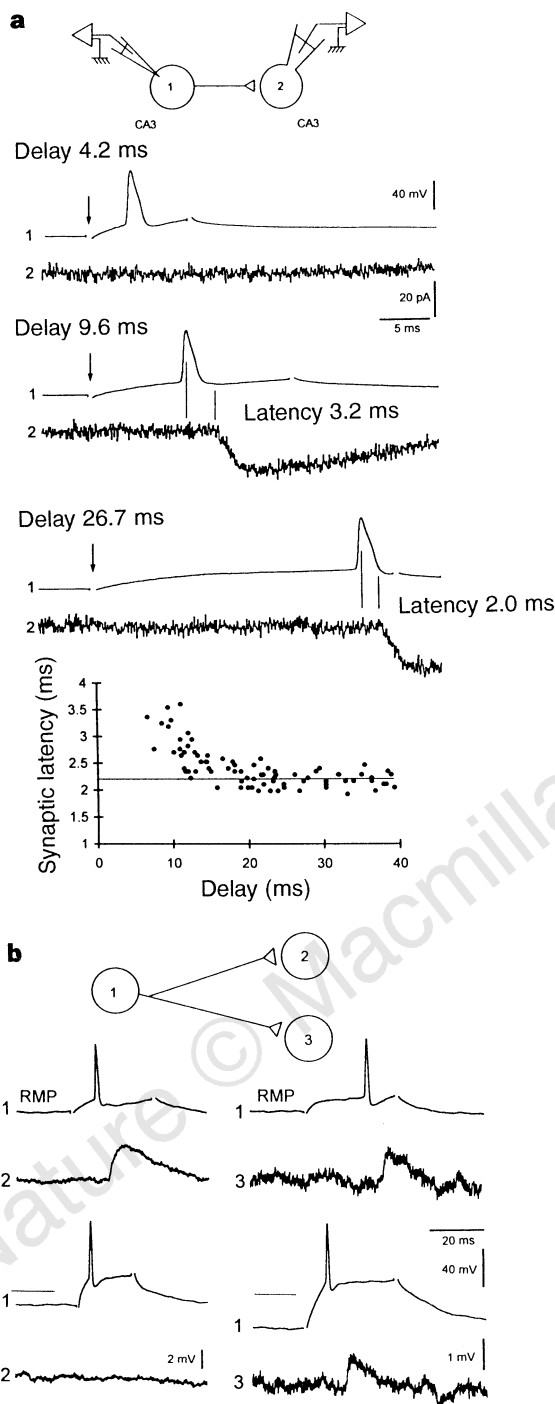


Figure 4 Propagation failures are associated with conduction delays and occur at branch points. **a**, Presynaptic neuron (1) hyperpolarized to -77 mV. A pAP induced with a 4.2 ms delay after onset of depolarizing pulse did not elicit an EPSC. Increasing delays permitted pAP propagation and reduced synaptic latency, as summarized underneath. Horizontal line indicates mean latency at resting membrane potential (RMP). **b**, When CA3 cell (1) was tonically hyperpolarized, no EPSPs were elicited in another CA3 cell (2). The electrode was then removed from cell (2) and inserted in CA3 cell (3). EPSPs were always generated. pAPs thus failed in the collateral of cell (1)'s axon to cell (2), but not in the branch to cell (3).

ing induction of an action potential in a neuron thus determines the functional coupling between cells. Unlike in invertebrates^{15,16} and dorsal root ganglion cells³, where pAPs fail only at frequencies > 10 Hz, propagation failures can also occur at low frequencies of stimulation. Furthermore, failures of a stimulus to elicit a post-synaptic response, usually interpreted as a transmitter release failure, can also result from failures of action-potential propagation²⁰.

In conclusion, axons need not always behave as simple cables, obliged to conduct action potentials faithfully from the cell body to nerve terminals (see also ref. 21): rather, we find that axons have the ability to 'integrate' recent changes in membrane potential—a facility previously reserved for neuronal dendrites and cell bodies. The recent past of the presynaptic neuron thus determines the fate of its output. □

Methods

Hippocampal slice cultures were prepared from 6-day-old rat pups and maintained for >2 weeks *in vitro* as described⁶. For electrophysiological recordings, cultures were transferred to a recording chamber mounted on an inverted microscope and continuously superfused with warmed (32°C) saline containing (in mM): Na^{+} , 149; Cl^{-} , 149; K^{+} , 2.7; Ca^{2+} , 2.8; Mg^{2+} , 2.0; HCO_3^{-} , 11.6; $\text{H}_2\text{PO}_4^{-}$, 0.4; glucose, 5.6; and phenol red (10 mg l^{-1}) at pH 7.4. Pyramidal neurons were impaled in stratum pyramidale using sharp microelectrodes filled with 1 M potassium methylsulphate⁵. In some experiments, postsynaptic cells were recorded with whole-cell techniques⁶. The analog signals from the two electrodes were digitized at 18 kHz and recorded on a video tape recorder. Off-line acquisition of 200-ms sequences was performed on an IBM PC (Acquis1, Bio-logic, Claix, France).

Received 13 March; accepted 16 June 1997.

1. Eccles, J. C. *The Physiology of Synapses* (Springer, Berlin, 1964).
2. Mackenzie, P. J., Umeyama, M. & Murphy, T. H. Ca^{2+} imaging of CNS axons in culture indicates reliable coupling between single action potentials and distal functional release sites. *Neuron* **16**, 783–795 (1996).
3. Lüscher, C., Streit, J., Quadroni, R. & Lüscher, H.-R. Action potential propagation through embryonic dorsal root ganglion cells in culture. I. Influence of the cell morphology on propagation properties. *J. Neurophysiol.* **72**, 622–633 (1994).
4. Debanne, D., Guérineau, N. C., Gähwiler, B. H. & Thompson, S. M. Physiology and pharmacology of unitary synaptic connections between pairs of cells in areas CA3 and CA1 of rat hippocampal slice cultures. *J. Neurophysiol.* **73**, 1282–1294 (1995).
5. Debanne, D., Guérineau, N. C., Gähwiler, B. H. & Thompson, S. M. Paired-pulse facilitation and depression at unitary synapses in rat hippocampus: quantal fluctuation affects subsequent release. *J. Physiol. (Lond.)* **491**, 163–176 (1996).
6. Gähwiler, B. H. Organotypic monolayer cultures of nervous tissue. *J. Neurosci. Meth.* **4**, 329–342 (1981).
7. Vincent, P. & Marty, A. Fluctuations in inhibitory postsynaptic currents in Purkinje cells from rat cerebellar slices. *J. Physiol. (Lond.)* **494**, 183–199 (1996).
8. Frotscher, M. & Gähwiler, B. H. Synaptic organization of intracellularly stained CA3 pyramidal neurons in slice cultures of rat hippocampus. *Neuroscience* **24**, 541–551 (1988).
9. Stuart, G. J. & Sakmann, B. Active propagation of somatic action potentials into neocortical pyramidal cell dendrites. *Nature* **367**, 69–72 (1994).
10. Colbert, C. M. & Johnston, D. Axonal action-potential initiation and Na^{+} channel densities in the soma and axon initial segment of subicular pyramidal neurons. *J. Neurosci.* **16**, 6676–6686 (1996).
11. Bossu, J.-L., Capogna, M., Debanne, D., McKinney, R. A. & Gähwiler, B. H. Somatic voltage-gated potassium currents of rat hippocampal pyramidal cells in organotypic slice cultures. *J. Physiol. (Lond.)* **495**, 367–381 (1996).
12. Stephens, G. J., Garratt, J. C., Robertson, B. & Owen, D. G. On the mechanism of 4-aminopyridine action on the cloned mouse brain potassium channel mKv1.1 . *J. Physiol. (Lond.)* **477**, 187–196 (1994).
13. Storm, J. F. Potassium currents in hippocampal pyramidal cells. *Progr. Brain Res.* **83**, 161–187 (1990).
14. Segue, I. Computer study of presynaptic inhibition controlling the spread of action potentials into axonal terminals. *J. Neurophysiol.* **63**, 987–998 (1990).
15. Parnas, I. Differential block at high frequency of branches of a single axon innervating two muscles. *J. Neurophysiol.* **35**, 903–914 (1972).
16. Parnas, I., Hochstein, S. & Parnas, H. Theoretical analysis of parameters leading to frequency modulation along an inhomogeneous axon. *J. Neurophysiol.* **39**, 909–923 (1976).
17. Li, X.-G., Somogyi, P., Ylinen, A. & Buzsáki, G. The hippocampal CA3 network: an *in vivo* intracellular labeling study. *J. Comp. Neurol.* **339**, 181–208 (1994).
18. Sheng, M., Tsaur, M.-L., Jan, Y. N. & Jan, L. Y. Subcellular segregation of two A-type K^{+} channel proteins in rat central neurons. *Neuron* **9**, 271–284 (1992).
19. Kocsis, J. D., Gordon, T. R. & Waxman, S. G. Mammalian optic nerve fibers display two pharmacologically distinct potassium channels. *Brain Res.* **383**, 357–361 (1986).
20. Storm, J. F. & Lipowsky, R. Evidence that excitability changes in presynaptic fibers may affect paired-pulse facilitation in hippocampal slices. *Soc. Neurosci. Abstr.* **20**, 1339 (1994).
21. Wall, P. D. Do nerve impulses penetrate terminal arborizations? A pre-synaptic control mechanism. *Trends Neurosci.* **18**, 99–103 (1995).

Acknowledgements. We thank L. Heeb and L. Rietschin for technical assistance, and A. Marty, J.-C. Ponce, B. Sakmann and S. Tyc-Dumont for critically reading the manuscript. Supported by the Dr. Eric Slack-Gyr and Swiss National Science Foundations. Parts of these data were presented to the American Society for Neuroscience in 1996.

Correspondence and requests for materials should be addressed to D.D. (e-mail: debanne@infobiogen.fr).

Phospholipase C isozymes selectively couple to specific neurotransmitter receptors

Daesoo Kim*, Ki Sun Jun*, Seong Beom Lee, Nae-Gyu Kang, Do Sik Min, Young-Hoon Kim†, Sung Ho Ryu, Pann-Ghill Suh & Hee-Sup Shin

Department of Life Science, Pohang University of Science and Technology, San-31, Hyoja-Dong, Pohang, 790-784, Republic of Korea

† Institute of Neuroscience, Inje University, 633-165, Gaegum-Dong, 614-735, Pusan, Republic of Korea

* These authors contributed equally to this work.

A variety of extracellular signals are transduced across the cell membrane by the enzyme phosphoinositide-specific phospholipase C- β (PLC- β) coupled with guanine-nucleotide-binding G proteins¹. There are four isoenzymes of PLC- β , β 1– β 4, but their functions *in vivo* are not known. Here we investigate the role of PLC- β 1 and PLC- β 4 in the brain by generating null mutations in mice: we found that PLC β 1^{−/−} mice developed epilepsy and PLC β 4^{−/−} mice showed ataxia. We determined the molecular basis of these phenotypes and show that PLC- β 1 is involved in signal transduction in the cerebral cortex and hippocampus by coupling predominantly to the muscarinic acetylcholine receptor, whereas PLC- β 4 works through the metabotropic glutamate receptor in the cerebellum, illustrating how PLC- β isoenzymes are used to generate different functions in the brain.

The mammalian phosphoinositide PLCs are grouped into three principal classes, β , γ and δ , on the basis of sequence homology and the location of structural domains². The PLC- γ family participates in the mitogenic response by direct coupling to both receptor- and non-receptor tyrosine kinase, whereas the β -family acts through G-protein-dependent pathways. However, it is not clear what roles the isoenzymes of a given PLC family play *in vivo*, or to what extent they determine the specificity of individual signalling pathways.

The expression patterns of PLC- β isoenzymes have not been well characterized except in the brain. In the brain, PLC- β 1, β 3 and β 4 are the prevalent forms and each isoenzyme has a unique distribution pattern. PLC- β 1 is highly expressed in the cerebral cortex and hippocampus³, and less so in the cerebellum; expression of PLC- β 4 is highest in the cerebellum and is almost negligible in the cerebral cortex and hippocampus⁴; and expression of PLC- β 3 is low throughout the brain⁴. To investigate how each isoenzyme functions *in vivo*, we selected PLC- β 1 and PLC- β 4 because their expression patterns are well defined in the brain. Deletion of PLC- β 4 by gene targeting to give PLC β 4^{−/−} mice⁵ has shown that PLC- β 4 functions in visual processing, although the molecular mechanism is unknown. We have now analysed the functions of PLC- β isozymes in brain by generating null mutants lacking either the PLC- β 1 or PLC- β 4 gene (Fig. 1a–d).

Mice heterozygous for the PLC- β 1 or the PLC- β 4 null mutation were normal and fertile, whereas both homozygotes displayed retarded growth and low viability after birth. Most PLC β 1^{−/−} mice died suddenly, starting at the third week after birth (Fig. 2a). The death of these mice was preceded by status epilepticus, or recurrent seizure attacks. The seizure was a generalized type characterized by tonic-clonic or tonic extension of the whole body (Fig. 1e) and resembled that induced by the GABA_A-receptor antagonist pentamethylenetetrazole (metrazol), blocking the inhibitory neuronal pathway. A tonic extension of the whole body is frequently observed in metrazol-induced seizures but rarely in seizures induced by kainic acid, a glutamate analogue that activates the excitatory neuronal pathway^{6,7}. This result suggests that PLC- β 1 is essential

for the normal functioning of inhibitory neuronal circuitry. PLC β 1^{−/−} mice were also shown to be hypersensitive to the epileptogenic drugs metrazol and kainic acid (Fig. 2b).

Histochemical analysis of PLC β 1^{−/−} hippocampus after spontaneous seizure revealed that only the somatostatin-containing interneurons were selectively lost in the hilus (Fig. 3A, b and f). The hilar interneurons containing somatostatin were also the ones selectively lost by electrical stimulation of adjacent granule cells of hippocampus⁸, indicating that seizures in PLC β 1^{−/−} mice were evoked through hyperexcitability of hippocampus. The loss of somatostatin-containing interneurons may cause the recurrent seizures associated with human temporal-lobe epilepsy⁹ and experimental seizure models⁸.

PLC β 4^{−/−} mice had a motor defect, broadly defined as ataxia. They were hypokinetic and showed a waddling gait, with the rear body swinging left–right. Footprint analysis of adult mice at postnatal day P27 revealed that mutant mice dragged and twisted their hind legs, with their toes spread apart (Fig. 1f). The ataxia was not due to malformation of bone structure or to muscle weakness: X-ray examination revealed that these mice had normal bone structure (data not shown) and a hanging test for muscle strength revealed no significant difference between wild-type and PLC β 4^{−/−} mice (duration of hanging on a rope: wild-type, 135.3 ± 14.9 s, *n* = 3; PLC β 4^{−/−}, 152.3 ± 15.9 s, *n* = 3; *P* = 0.546). The motor defect could account for the growth retardation and postnatal lethality of PLC β 4^{−/−} mice because they were at a disadvantage when competing with normal littermates: the mutants survived and grew well in the absence of normal mice.

Histological analysis of PLC β 4^{−/−} mice revealed that the development of the cerebellum was retarded, being smaller in size, with aberrant patterns of folia and incompletely migrated external granule cells (Fig. 3B, d and f) at P15. At P50, the mutant cerebellum was apparently normal except for a slight reduction in size (Fig. 3B, a and b), although the mice were still ataxic. Dendrogenesis of Purkinje cells examined by anti-calbindin staining and the number of synapses on dendritic spines, as examined by transmission electron microscopy, were also normal in the cerebellum of PLC β 4^{−/−} mice at 50 (data not shown), suggesting that the motor defect in these mice was not due to any grossly abnormal development of the cerebellum. As shown below, the defect was, at least in part, a result of impaired PLC-linked signal transduction.

To determine the molecular basis of the phenotypic defects of the mutant mice, we tried to identify the cognate receptors that are coupled to each isozyme by examining specific agonist-stimulated phosphoinositide turnover in the brain. We selected three agonists on the basis of the known expression patterns of corresponding PLC-coupled receptors: carbachol (CCh), a non-selective agonist on muscarinic acetylcholine receptors (mAChR); 1-amino-1,3-cyclopentanedicarboxylic acid (ACPD), a non-selective agonist on metabotropic glutamate receptors (mGluR); and 1-(3-chlorophenyl)piperazine (mCPP), a non-selective agonist on serotonin (5-HT) type-2 receptors. None of these receptors has so far been found to couple to a specific PLC- β isoenzyme *in vivo*.

Figure 4a shows that, in the PLC β 1^{−/−} hippocampus, CCh-induced phosphoinositide hydrolysis was markedly attenuated, whereas ACPD-induced phosphoinositide hydrolysis was increased significantly. mCPP-induced phosphoinositide hydrolysis did not change at all. These results indicate that PLC- β 1 deficiency has selectively impaired mACh signalling in the hippocampus. The attenuation of mACh signalling was also apparent in other PLC β 1-expressing regions of the brain, temporal cortex and cerebellum (Fig. 4a). Our conclusion is consistent with previous findings that the expression pattern of the M1 and M3 AChR¹⁰ overlaps with that of PLC- β 1 in various regions of the brain² and with those from reconstitution experiments in which the M1 receptor was shown to couple *in vitro* with purified PLC- β 1 via G protein¹¹.

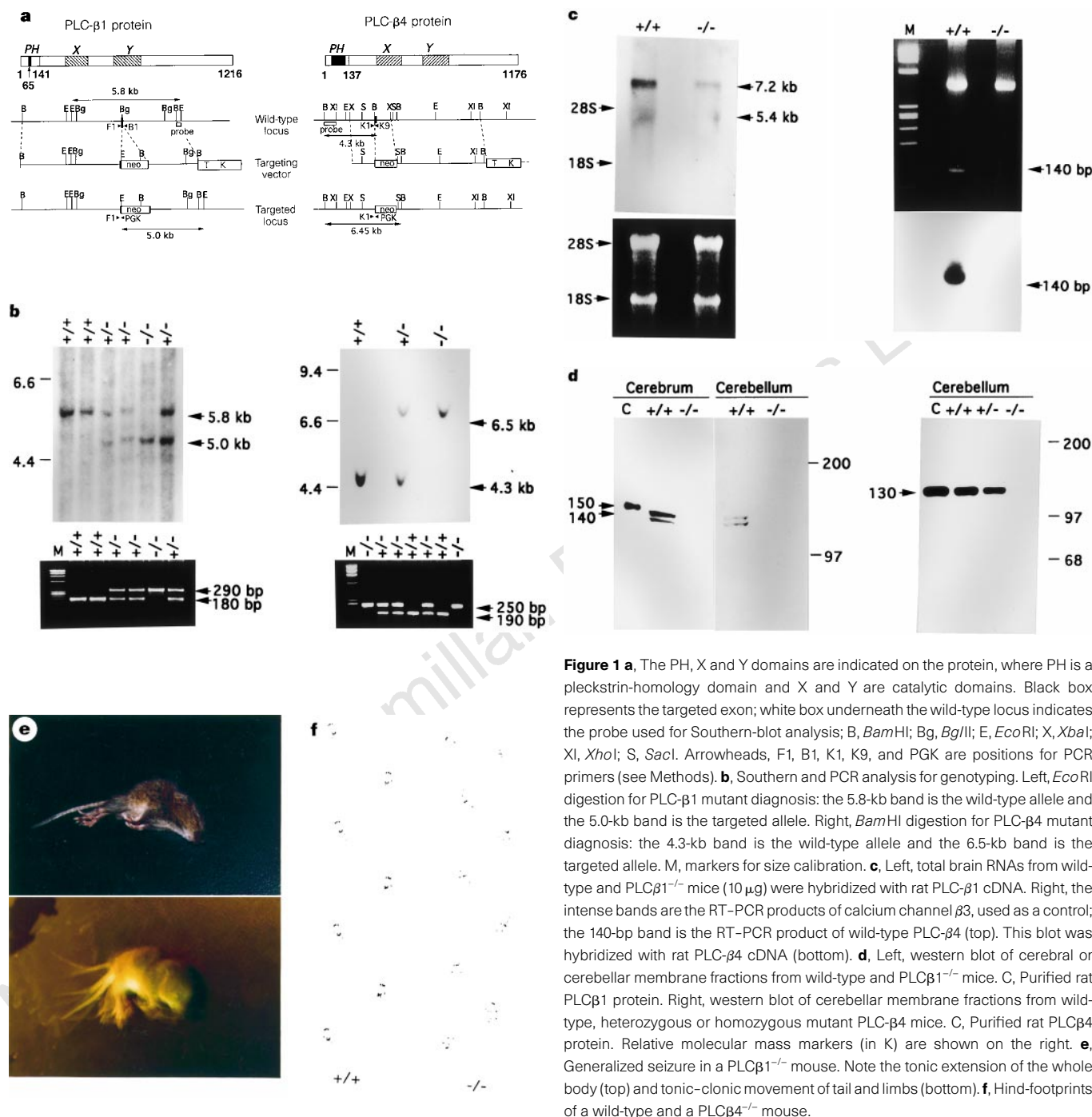


Figure 1 **a**, The PH, X and Y domains are indicated on the protein, where PH is a pleckstrin-homology domain and X and Y are catalytic domains. Black box represents the targeted exon; white box underneath the wild-type locus indicates the probe used for Southern-blot analysis; B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; X, *Xba*I; XI, *Xho*I; S, *Sac*I. Arrowheads, F1, B1, K1, K9, and PGK are positions for PCR primers (see Methods). **b**, Southern and PCR analysis for genotyping. Left, *Eco*RI digestion for PLC-β1 mutant diagnosis: the 5.8-kb band is the wild-type allele and the 5.0-kb band is the targeted allele. Right, *Bam*HI digestion for PLC-β4 mutant diagnosis: the 4.3-kb band is the wild-type allele and the 6.5-kb band is the targeted allele. M, markers for size calibration. **c**, Left, total brain RNAs from wild-type and PLCβ1^{-/-} mice (10 μg) were hybridized with rat PLC-β1 cDNA. Right, the intense bands are the RT-PCR products of calcium channel β3, used as a control; the 140-bp band is the RT-PCR product of wild-type PLC-β4 (top). This blot was hybridized with rat PLC-β4 cDNA (bottom). **d**, Left, western blot of cerebral or cerebellar membrane fractions from wild-type and PLCβ1^{-/-} mice. C, Purified rat PLCβ1 protein. Right, western blot of cerebellar membrane fractions from wild-type, heterozygous or homozygous mutant PLC-β4 mice. C, Purified rat PLCβ4 protein. Relative molecular mass markers (in K) are shown on the right. **e**, Generalized seizure in a PLCβ1^{-/-} mouse. Note the tonic extension of the whole body (top) and tonic-clonic movement of tail and limbs (bottom). **f**, Hind-footprints of a wild-type and a PLCβ4^{-/-} mouse.

Because null mutants of the 5-HT_{2c} receptor also suffer from epilepsy⁶, we suspected that PLC-β1 might be a transducer of 5-HT_{2c} signalling in the hippocampus. PLC-β1 deficiency, however, was not associated with significant loss of 5-HT_{2c}R linked phosphoinositide hydrolysis (Fig. 4a), indicating that the molecular mechanism of PLCβ1^{-/-} epilepsy may differ from that in 5-HT_{2c}R^{-/-} mice. As the axotomy of a cholinergic input, septo-hippocampal pathway triggers chronic seizures in the hippocampus¹² and cholinergic stimulation in hippocampal slices decreases excitability by activating inhibitory interneurons¹³, the loss of cholinergic signalling in the PLCβ1^{-/-} hippocampus may explain the seizure activity in PLCβ1^{-/-} mice. This kind of seizure activity could lead to the loss of somatostatin-containing inter-

neurons, as already mentioned, and thereby establish persistent epileptic states in the PLCβ1^{-/-} hippocampus.

The alteration of cholinergic signal transduction in the hippocampus may be an important pathogenic process in several neurological disorders, including dementia and Alzheimer's disease¹⁴. The cortex of Alzheimer's patients is insensitive to carbachol-induced phosphoinositide hydrolysis¹⁵ because β-amyloid may disturb mAChR-PLC coupling¹⁶. The pathological significance of lost mAChR signalling is unclear, but our data indicate that the cholinergic defects may account, at least in part, for the increased susceptibility to seizure in patients with Alzheimer's^{17,18}.

In cerebellum, PLC-β1 and β4 are both expressed but at different levels²³: PLC-β4 is more abundant than PLC-β1 in normal cerebellum

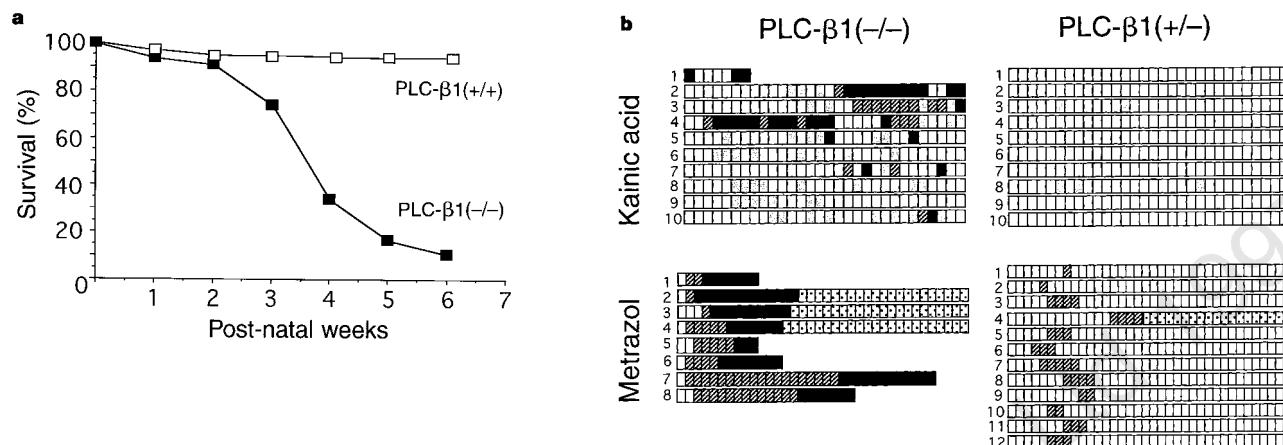


Figure 2 Epileptic seizures in PLCβ1^{-/-} mice. **a**, Analysis of survival rate of wild-type (white squares) and PLCβ1^{-/-} (black squares) mice ($n = 36$). **b**, Convulsive behaviour after administration of kainate or metrazol at subconvulsive doses. Black boxes, generalized seizures showing clonic, tonic-clonic or tonic extension of the whole body; cross-hatched, tonic activities such as temporal tonic exten-

sion of limb and tail, tremor, or absence; shaded, stereotypy such as repetitive scratchings, circlings or rearings; square-patterned, post-ictal periods composed of hypokinesia, sleeping or recovery of body control. Timescale in the kainate experiment was one box, 2 min; in the metrazol experiment it was a white or cross-hatched box, 20 s; a black or a dotted box, 200 s.

(Fig. 1d). As shown in Fig. 4b, in the cerebellum of PLCβ4^{-/-} mice there is less phosphoinositide hydrolysis induced by CCh or ACPD, whereas PLCβ1^{-/-} mice show a preferential attenuation of CCh-induced phosphoinositide hydrolysis as in hippocampus. Thus, in cerebellum, PLC-β4 appears to be the transducer of signals mediated by mGluR, whereas PLC-β1 and β4 both transduce mAChR-mediated signals. As PLC-β4 is enriched in Purkinje cells³, the ataxia phenotype of PLCβ4^{-/-} mice may, at least in part, be the result of an impaired mGluR signal-transduction pathway in Purkinje cells. This is supported by the findings that mGluR1 is expressed predominantly in cerebellar Purkinje cells¹⁹ and that

mGluR1 knockout mice show ataxia²⁰. Thus, PLC-β4 appears to be essential for transducing mGluR1-mediated signals in cerebellum, which accounts for the defect seen in PLCβ4^{-/-} mice.

Our results show that the PLC-β isoenzymes β1 and β4 have different functions in the brain. Structural differences as well may account for the functional divergence of the two isoenzymes. For example, the sequence homology of PLC-β isoenzymes is minimal in the carboxy-terminal region that interacts with the G-protein complex². The M1 muscarinic receptor interacts only with G-protein complexes composed of the subunits αq, α11, β1, β4 and γ4 in RBL-2H3 cells, despite the fact that other subunits are also

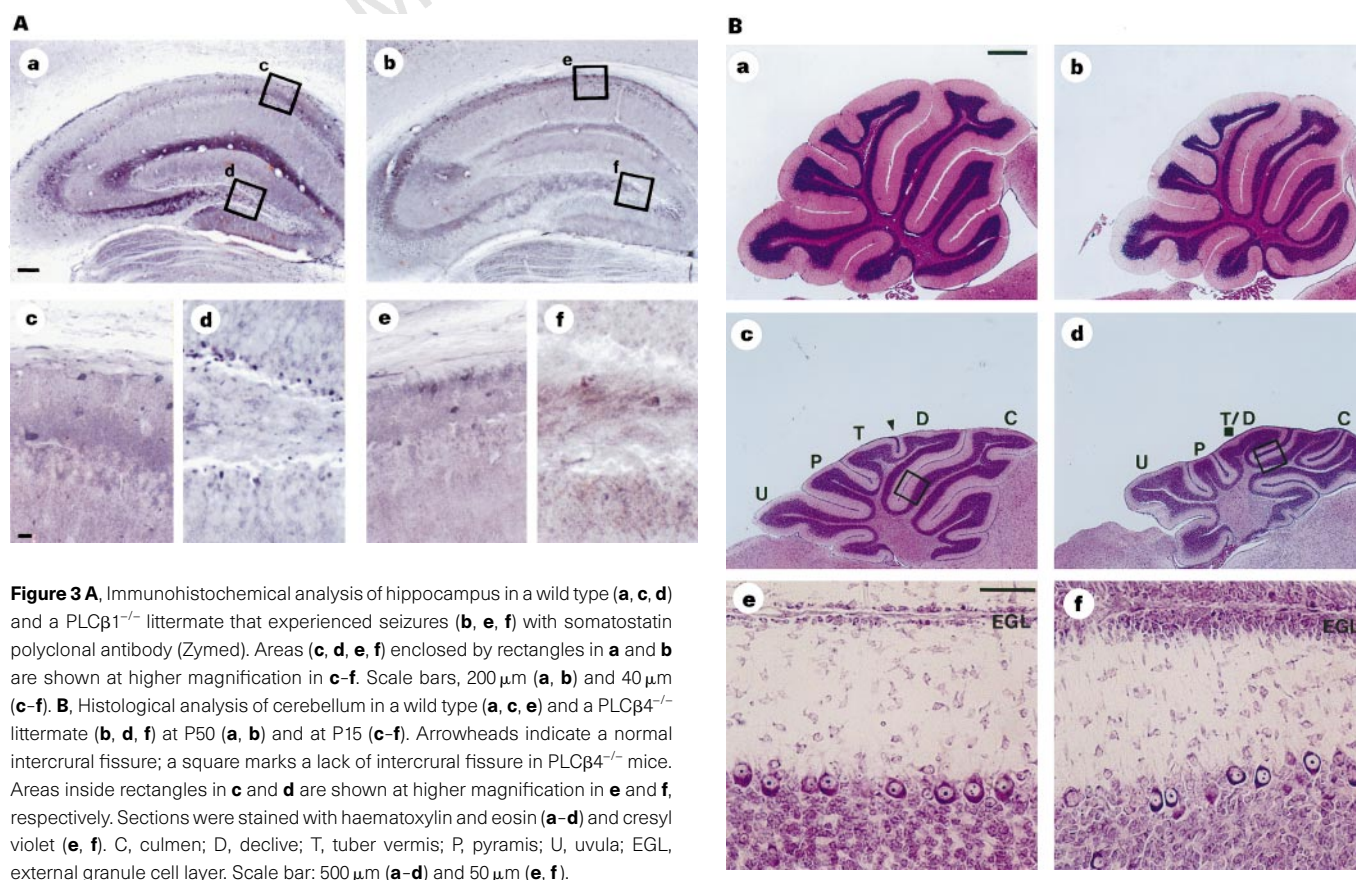


Figure 3 A, Immunohistochemical analysis of hippocampus in a wild type (**a, c, d**) and a PLCβ1^{-/-} littermate that experienced seizures (**b, e, f**) with somatostatin polyclonal antibody (Zymed). Areas (**c, d, e, f**) enclosed by rectangles in **a** and **b** are shown at higher magnification in **c-f**. Scale bars, 200 μm (**a, b**) and 40 μm (**c-f**). **B**, Histological analysis of cerebellum in a wild type (**a, c, e**) and a PLCβ4^{-/-} littermate (**b, d, f**) at P50 (**a, b**) and at P15 (**c-f**). Arrowheads indicate a normal intercrural fissure; a square marks a lack of intercrural fissure in PLCβ4^{-/-} mice. Areas inside rectangles in **c** and **d** are shown at higher magnification in **e** and **f**, respectively. Sections were stained with haematoxylin and eosin (**a-d**) and cresyl violet (**e, f**). C, culmen; D, declive; T, tuber vermis; P, pyramis; U, uvula; EGL, external granule cell layer. Scale bar: 500 μm (**a-d**) and 50 μm (**e, f**).

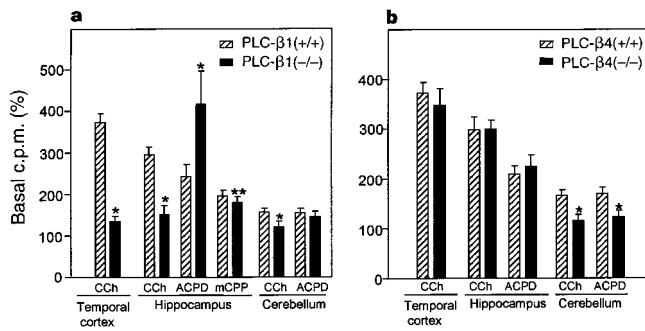


Figure 4 LiCl-amplified agonist-induced phosphoinositide (PI) hydrolysis in **a**, PLC-β1^{-/-}, and **b**, PLC-β4^{-/-} brain slices. Agonist-induced PI hydrolysis was assayed as described in Methods. Treatments with CCh (1 mM), ACPD (100 μM) or mCPP (2 mM) were performed in triplicate three times. Basal c.p.m. of 20 μl brain slices (the volume for one reaction) varied between different brain regions but not between mutants and wild type, being 1,535 ± 237 for temporal cortex, 1,014 ± 181 for hippocampus, and 1,852 ± 413 for cerebellum. Student's *t*-test: **P* < 0.01; ***P* = 0.06.

expressed in the cell²¹. Thus, a specific combination of G proteins active in signalling from a given receptor may be involved in the selective activation of PLC-β isoenzymes.

Finally, our results suggest a pharmacological basis for the neurological disorders displayed by the PLC-β1-deficient mice and β4. These knockout mice may be able to model similar human disorders, for example PLC-β1^{-/-} complement EL2, a previously identified epilepsy locus²² (data not shown) so providing a new epilepsy model. PLC-β1^{-/-} mice may also be useful for analysing the pathogenic role of muscarinic cholinergic loss in Alzheimer's patients. □

Methods

Gene targeting. The 7.6-kb *Bam*HI fragment containing the exon that encodes amino-acid residues 50–82 of PLC-β1 was subcloned from a genomic clone hybridized with a 0.2-kb *Bgl*II–*Acc*I fragment of rat PLC-β1 cDNA (gift from S. G. Rhee). The PGK-neo cassette from pWH11 and SV40 poly(A) from pWH17 (ref. 23) was inserted into the *Bgl*III site of the exon for gene disruption and positive selection. The thymidine kinase (TK) gene cassette was attached to the end of the 3' homology region for negative selection. The vector for PLC-β4 targeting was designed to disrupt the brain-specific isoform²⁴. A genomic clone having an exon that covers the nucleotide sequence 226–369 from the initiation codon of cDNA was isolated by using the 2.75-kb *Sac*I–*Eco*RI fragment of rat brain PLC-β4 cDNA as probe. The exon, together with a portion of the 3' flanking intron sequence was replaced with the *neo* cassette. The TK cassette was attached to the end of the 3' homology region of the targeting vector.

Genotyping by PCR. Genomic DNA was prepared from tail biopsies. For PLC-β1 mutants, F1(5'-CAAGTTAAGTCCGCAACACC-3'), B1(5'-ACCTTG-GG AGCTTTGGCGTG-3') and PGK22(5'-CTGACTAGGGGAGGAGTAGA-AG-3') primers were used. For PLC-β4 mutants, K1(5'-CTCCACACTGTCAAC CTAC-3'), K9(5'-AGTACTTCTGGATTTCAGCC-3') and PGK22 were used.

Drug-induced epilepsy. Mice were weaned at age P21 and housed with their siblings in a cage. At age P21–26, mice were injected intraperitoneally with 10 mg kg⁻¹ kainate or 50 mg kg⁻¹ metrazol. They were then placed in a plastic cage and their behaviour recorded by a video camera.

Histological analysis. Mice were perfused transcardially with 4% paraformaldehyde under anaesthesia. Brains were embedded in paraffin wax, and serial sagittal or coronal sections (7 μm) mounted on slides and stained with haematoxylin and eosin or cresyl violet. For immunohistochemistry, 50-μm brain sections containing the hippocampal structure were prepared using a vibratome, stained with primary antibodies and visualized by using a peroxidase kit (Vector).

Phosphoinositide hydrolysis in brain slices. Chopped brain slices (hippocampus at 0.35 mm, temporal cortex at 0.35 × 0.35 mm, and cerebellum at 0.4 × 0.35 mm) were used for phosphoinositide hydrolysis experiments as described²⁵, with some modification. Briefly, 20 μl of gravity-packed brain slices were prelabelled with 2-³[H]-myo-inositol (5 μCi ml⁻¹) in 0.25 ml Krebs' bicarbonate buffer (containing 1.3 mM CaCl₂ and saturated with 95% O₂/5% CO₂) for 1 h, treated with LiCl (10 mM) and myo-inositol (10 mM) for 20 min, and then incubated with or without an appropriate agonist for 40 min. All procedures were performed in capped 5-ml flat-bottomed tubes (Amicon) under periodical supplies of gas at 37 °C. The reaction was stopped with 1 ml CHCl₃/CH₃OH (1:2) and extracted with 0.6 ml CHCl₃/H₂O (1:1). Total inositol polyphosphates in the aqueous phase were isolated on an AG1X-8 column. Agonist-induced phosphoinositide hydrolysis was expressed as a percentage of basal c.p.m. as described²⁶ (basal c.p.m. were obtained in a reaction without agonist, the result of the action of endogenous agonists in the slices²⁵).

Received 18 July; accepted 5 August 1997.

- Fisher, S. K., Heacock, A. M. & Agarano, B. W. Inositol lipids and signal transduction in the nervous system: An update. *J. Neurochem.* **58**, 18–38 (1992).
- Rhee, S. G. & Bae, Y. S. Regulation of phosphoinositide specific phospholipase C isozymes. *J. Biol. Chem.* **272**, 15045–15048 (1997).
- Ross, C. A., MacCumber, M. W., Glatt, C. E. & Snyder, S. H. Brain phospholipase C isozymes: Differential mRNA localizations by *in situ* hybridization. *Proc. Natl Acad. Sci. USA* **86**, 2923–2927 (1989).
- Tanaka, O. & Kondo, H. Localization of mRNA for three novel members (β3, β4 and γ2) of phospholipase C family in mature rat brain. *Neurosci. Lett.* **182**, 17–20 (1994).
- Jiang, H. et al. Phospholipase C β4 is involved in modulating the visual response in mice. *Proc. Natl Acad. Sci. USA* **93**, 14598–14601 (1996).
- Tecott, L. H. et al. Eating disorder and epilepsy in mice lacking 5-HT_{2c} serotonin receptors. *Nature* **374**, 542–546 (1995).
- Stafstrom, C. E., Thompson, J. L. & Holms, G. L. Kainic acid seizures in the developing brain: Status epilepticus and spontaneous recurrent seizures. *Dev. Brain Res.* **65**, 227–236 (1992).
- Sloviter, R. S. Decreased hippocampal inhibition and a selective loss of interneurons in experimental epilepsy. *Science* **235**, 73–76 (1987).
- de Lanerolle, N. C., Kim, J. H., Robbins, R. J. & Spencer, D. D. Hippocampal interneuron loss and plasticity in human temporal lobe epilepsy. *Brain Res.* **495**, 387–395 (1989).
- Buckley, N. J., Bonner, T. I. & Brann, M. R. Localization of muscarinic receptor mRNAs in rat brain. *J. Neurosci.* **8**, 4646–4652 (1988).
- Biddlecome, G. H., Bernstein, G. & Ross, E. M. Regulation of phospholipase C-β1 by Gq and m1 muscarinic cholinergic receptor. Steady-state balance of receptor-mediated activation and GTPase-activating protein-promoted deactivation. *J. Biol. Chem.* **271**, 7999–8007 (1996).
- Buzsaki, G., Ponomareff, G. L., Bayardo, F., Ruiz, R. & Gage, F. H. Neuronal activity in the subcortically denervated hippocampus: a chronic model for epilepsy. *Neuroscience* **28**, 527–538 (1989).
- Pitler, T. A. & Alger, B. E. Cholinergic excitation of GABAergic interneurons in the rat hippocampal slice. *J. Physiol. (Lond.)* **450**, 127–142 (1992).
- Bartus, R. T., Dean, R. I. III, Beer, B. & Lippa, A. S. The cholinergic hypothesis of geriatric memory dysfunction. *Science* **217**, 408–414 (1982).
- Crews, F. T., Kurian, P. & Freund, G. Cholinergic and serotonergic stimulation of phosphoinositide hydrolysis is decreased in Alzheimer disease. *Life Sci.* **55**, 1993–2002 (1994).
- Kelly, J. F. et al. Amyloid β-peptide disrupts carbachol-induced muscarinic cholinergic signal transduction in cortical neurons. *Proc. Natl Acad. Sci. USA* **93**, 6753–6758 (1996).
- Hesdorffer, D. C., Hauser, W. A., Annegers, J. F., Kokmen, E. & Rocca, W. A. Dementia and adult-onset unprovoked seizures. *Neurology* **46**, 727–730 (1996).
- Mendez, M. F., Catanzaro, P., Doss, R. C., Arguello, R. & Frey, W. H. Seizures in Alzheimer's disease: clinicopathologic study. *J. Geriatr. Psychiatr. Neurol.* **7**, 230–233 (1994).
- Masu, M., Tanabe, Y., Tsuchida, K., Shigemoto, R. & Nakanishi, S. Sequence and expression of a metabotropic glutamate receptor. *Nature* **349**, 760–765 (1991).
- Aiba, A. et al. Deficient cerebellar long-term depression and impaired motor learning in mGluR1 mutant mice. *Cell* **79**, 377–388 (1994).
- Dippel, E., Kalkbrenner, F., Wittig, B. & Schultz, G. A heterotrimeric G protein complex couples the muscarinic m1 receptor to phospholipase C-β. *Proc. Natl Acad. Sci. USA* **93**, 1391–1396 (1996).
- Frankel, W. N., Johanson, C. M. & Lutz, C. M. Congenic strains reveal effects of the epilepsy quantitative trait locus, EL2, separate from other EL loci. *Mamm. Genome* **6**, 839–843 (1995).
- Kim, D. G., Kang, H. M., Jang, S. K. & Shin, H.-S. Construction of a bifunctional mRNA in the mouse by using the internal ribosomal entry site of the encephalomyocarditis virus. *Mol. Cell. Biol.* **13**, 3636–3643 (1992).
- Kim, M. J. et al. Cloning of cDNA encoding phospholipase C-β4, a new member of the phospholipase C. *Biochem. Biophys. Res. Commun.* **194**, 706–712 (1993).
- Berridge, M. J. et al. Change in the levels of inositol phosphates after agonist-dependent hydrolysis of membrane phosphoinositides. *Biochem. J.* **212**, 473–482 (1983).
- Schoepp, D. D., Johnson, B. G., True, R. A. & Monn, J. A. Comparison of (1S,3R)-1-aminocyclopentane-1,3-dicarboxylic acid (1S,3R-ACPD) and 1R,3S-ACPD-stimulated brain phosphoinositide hydrolysis. *Eur. J. Pharmacol.* **207**, 351–353 (1991).

Acknowledgements. This Letter is dedicated to the late Dr. Hogil Kim, the former president of POSTECH, who made this work possible. H.-S.S. thanks E. Park for comments on the manuscript and E. B. Lee and B. J. Pak for their early contribution to this work. We thank S. G. Rhee for PLC-β1 cDNA, Y. Namkung for ES cell culture, I. S. Kim, I. B. Kim, H. D. Noh, C. T. Choi and I. S. Lee for help with histology, and H. Y. Lee for video processing. This work was supported in part by a genetic engineering grant from the Ministry of Education and a medical science grant from the Ministry of Health and Welfare of Korea, and by grants from the Basic Science Research Institute and from the School of Environmental Engineering, POSTECH.

Correspondence and requests for materials should be addressed to H.-S.S. (e-mail:shinhs@vision.postech.ac.kr).

Stimulation of CFTR activity by its phosphorylated R domain

Michael C. Winter & Michael J. Welsh

Howard Hughes Medical Institute, Departments of Internal Medicine and Physiology and Biophysics, University of Iowa College of Medicine, Iowa City, Iowa 52242, USA

Phosphorylation controls the activity of ion channels in many tissues. In epithelia, the cystic fibrosis transmembrane conductance regulator (CFTR) chloride channel is activated by phosphorylation of serine residues in its regulatory (R) domain and then gated by binding and hydrolysis of ATP by the nucleotide-binding domains^{1–3}. Current models propose that the unphosphorylated R domain serves as an inhibitory particle that occludes the pore^{1,2,4–6}, much like the inhibitory ‘ball’ in Shaker K⁺ channels^{7,8}; presumably, phosphorylation relieves this inhibition. Here we test this by adding an R-domain peptide to a CFTR variant in which much of the R domain had been deleted (CFTR-ΔR/S660A): in contrast to predictions, we found that adding an unphosphorylated R domain to CFTR-ΔR/S660A did not inhibit activity, whereas a phosphorylated R-domain peptide stimulated activity. To investigate how phosphorylation controls activity, we studied channel gating and found that phosphorylation of the R domain increases the rate of channel opening by enhancing the sensitivity to ATP. Our results indicate that CFTR is regulated by a new mechanism in which phosphorylation of one domain stimulates the interaction of ATP with another domain, thereby increasing activity.

We previously studied a CFTR variant, CFTR-ΔR/S660A, in which much of the R domain (residues 708–835) was deleted and the phosphorylation site was removed at Ser 660 by mutation to alanine⁴. This variant does not need to be phosphorylated in order to open in the presence of ATP, and addition of the cyclic-AMP-dependent protein kinase A (PKA) does not alter its open-state probability (P_o). Based on those and other considerations, it has been proposed that the R domain inhibits the channel^{1–2,4–6,9}. This model predicts that the activity of CFTR-ΔR/S660A should be similar to that of phosphorylated wild-type CFTR. However, when we examined the single-channel properties of CFTR-ΔR/S660A, we found that P_o was 0.13 ± 0.01 (1 mM ATP, $n = 7$), much lower than that of phosphorylated wild-type CFTR ($P_o = 0.46 \pm 0.02$; 1 mM ATP, $n = 17$, $P < 0.001$). These results, which are similar to those described in a previous report⁴, indicated that the R domain might be more than just a plug that inhibits the channel. But deletion of the R domain in CFTR-ΔR/S660A could still have additional nonspecific effects on channel structure which might reduce activity. We considered, therefore, that the phosphorylated R domain might stimulate the channel and that the defective functioning of CFTR-ΔR/S660A is due to its lacking a phosphorylated R domain.

To test this idea, we added a recombinant R domain (R1, residues 645–834, similar to a peptide reported previously¹⁰) to the cytosolic surface of CFTR-ΔR/S660A and measured activity by using the excised, inside-out patch-clamp technique. Figure 1a, c shows that phosphorylated R1 stimulates the channel: that is, the current increases when the R domain is added in the presence of PKA (PKA phosphorylates R1 and related peptides¹⁰; S. Travis, unpublished results). Stimulation was reversed when R1 was removed. Figure 1b, c shows that R1 has no effect in the absence of PKA, indicating that it is the phosphorylated R domain that stimulates channel activity. Additional evidence that the effect of phosphorylation was on R1 rather than on CFTR-ΔR/S660A came from our

finding that addition of phosphorylated R1 in the presence of protein-kinase-inhibitor peptide under conditions that block the effect of PKA¹¹ gave similar results ($43.2 \pm 3.4\%$ increase; $n = 3$). A different PKA substrate (‘Kemptide’; 10 μM (ref. 12)) had no effect in the presence or absence of PKA ($n = 4$; results not shown). R1 had no effect on wild-type CFTR (Fig. 1c), perhaps because the endogenous R domain has a maximal effect or because steric hindrance from the endogenous R domain prevents interaction with R1. This latter possibility is consistent with our finding that phosphorylated R1 did not stimulate CFTR containing mutated phosphorylation sites ($n = 3$; not shown). Our results also suggest that the low P_o found for CFTR-ΔR/S660A results at least in part from a lack of stimulation by a phosphorylated R domain, rather than from nonspecific structural changes. The fact that R1 did not increase the P_o of CFTR-ΔR/S660A to wild-type CFTR values may be because the endogenous R domain is constrained in the correct orientation by the site of interaction; the interaction of exogenously applied R1 with CFTR-ΔR/S660A is probably less efficient. Our results contrast with a report that addition of an *in vitro* translation mixture containing a different unphosphorylated R domain (residues

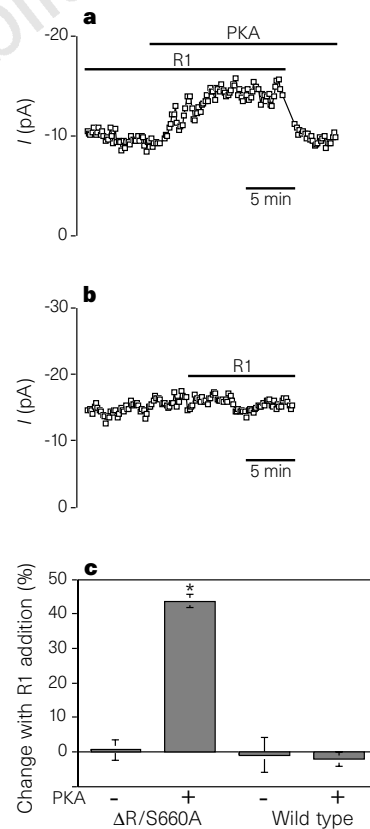


Figure 1 Effect of R1 on CFTR-ΔR/S660A and wild-type CFTR. **a, b**, Time course of Cl⁻ current in excised, inside-out patches containing many CFTR-ΔR/S660A channels during addition of PKA and R1 (50 nM), as indicated by bars. Each data point represents mean current during a 10 s interval. ATP (1 mM) was present at all times. Similar results were obtained irrespective of the sequence of R1 and PKA addition. **c**, Effect of R1 in the presence and absence of PKA on current measured from CFTR-ΔR/S660A and wild-type CFTR. Wild-type CFTR channels were initially phosphorylated with PKA and ATP. Then R1 was added in either the continued presence of PKA or following its removal, as indicated. Note that once they are phosphorylated, wild-type channels remain active in the absence of PKA (ref. 11, and see Fig. 3b). For CFTR-ΔR/S660A, n is 10 and 9; for wild-type CFTR, n is 4 and 5, in the absence or presence of PKA, respectively. Asterisk indicates $P < 0.01$ compared to value in the absence of R1. Values are mean current measured between 10 and 12 min after addition of R1.

588–858) inhibited wild-type CFTR channels in lipid bilayers⁵, but they do not preclude such effects under certain conditions.

To find out how the phosphorylated R domain stimulates CFTR- Δ R/S660A, we investigated channel gating. Gating of CFTR can be described by the rate of channel opening into bursts of activity and the rate of closing from bursts^{13–15}. Figure 2 shows that phosphorylated R1 increases the rate at which CFTR- Δ R/S660A opened into bursts, without affecting the rate of closure from bursts. These results predict that a decrease in phosphorylation of intact, wild-type CFTR would have the opposite effect, decreasing the rate of opening. We therefore tested the gating of CFTR variants in which endogenous R-domain serines were mutated to alanine. We studied four serine residues (Ser 660, 737, 795, 813) that are phosphorylated *in vivo*^{10,16,17}. Mutation of the individual serines reduced P_o , but to different extents: the S795A mutation had the largest effect and S737A had little or no effect in the presence of 1 mM ATP (Fig. 3a, b). Simultaneous mutation of all four serines (S-quad-A) also decreased P_o . Moreover, when we removed PKA from wild-type CFTR, P_o decreased, consistent with a previous suggestion that there is some phosphatase activity associated with the patch of membrane^{18–20}. The serine mutations decreased the rate of channel opening without altering the rate of closing (Fig. 3c, d).

Thus addition of the phosphorylated R domain to CFTR- Δ R/S660A and mutation of phosphorylation sites in wild-type CFTR had opposite effects on the same kinetic step in channel gating, the rate of channel opening. It has been shown that the rate of channel opening is controlled by the interaction of ATP with the nucleotide-binding domains (NBDs)^{13,20,21}. Therefore, our data indicated that the R domain may alter ATP-dependent and hence NBD-dependent regulation.

To test this idea, we examined the effect of ATP concentration on CFTR containing serine mutations in the R domain. Figure 4 shows that the mutations reduced P_o at low concentrations of ATP; P_o was

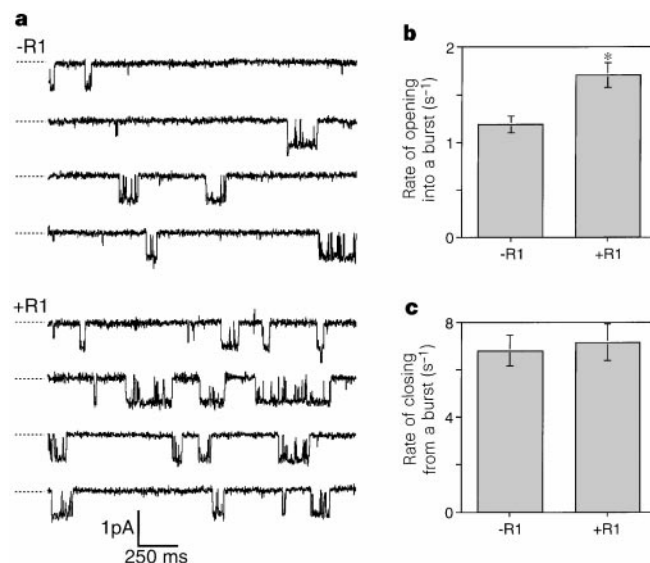


Figure 2 Effect of R1 on CFTR- Δ R/S660A channels in the presence of PKA. **a**, Example of continuous single-channel tracings before and at 10 min after the addition of R1. **b**, **c**, Effect of R1 on the rates of entry into and exit from a burst for CFTR- Δ R/S660A channels in the presence of PKA. Note that because they do not measure the fast closings during a burst, calculation of P_o from the rates in **b** and **c** gives an overestimate of P_o . Use of a 50% threshold-crossing method to determine P_o gives values of 0.134 ± 0.015 and 0.189 ± 0.021 in the absence and presence of R1, respectively. Data were obtained between 10 and 12 min after adding R1. PKA was present under all conditions; $n = 10$ in the absence and $n = 6$ in the presence of R1. The asterisk indicates difference from that in the absence of R1 ($P < 0.05$).

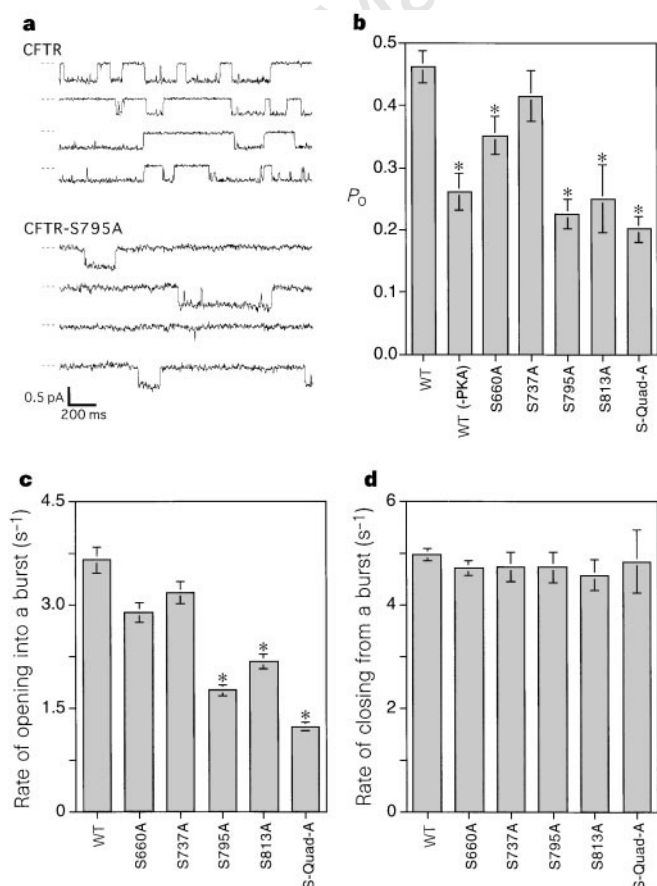


Figure 3 Effect of phosphorylation-site mutations on CFTR Cl⁻ channel activity. **a**, Example of continuous single-channel tracings from wild-type CFTR and CFTR-S795A; dashed lines indicate the closed state. **b**, **c** and **d**, P_o , opening and closing rates. PKA was present in **a–d**, except where noted. For studies in the absence of PKA (–PKA), the channel was first phosphorylated and then PKA was removed. ATP (1 mM) was present in all cases. Data were obtained using at least 10 min of continuous recording from each patch. Number of experiments for **b/c** and **d** were: 17/6 for wild-type (WT), 8 for wild-type (–PKA), 8/7 for S660A, 5/5 for S737A, 8/6 for S795A, 9/7 for S813A, and 11/4 for S-Quad-A. Data are mean $\pm$ s.e.m. Asterisk indicates values significantly different from wild type. There were no significant differences in measured burst duration for any of the mutants.

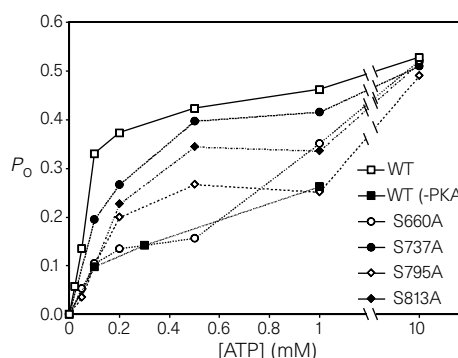


Figure 4 Effect of ATP concentration on P_o of CFTR containing phosphorylation site mutations. Data points are mean of at least 4 experiments; error bars are omitted for clarity (Fig. 3b shows error at 1 mM ATP).

also reduced for wild-type CFTR in the absence of PKA. These effects could be explained either by reduced ATP binding or by a reduction in some other step in the ATP hydrolysis cycle²⁰. However all mutants had the same P_o at 10 mM ATP, showing that the reduced activity could be overcome by using high ATP concentrations. This might also be true for CFTR- $\Delta R/S660A$. Because all the steps in the ATP hydrolysis cycle are independent of ATP concentration apart from ATP binding, our results suggest that the mutations reduce ATP binding. We cannot distinguish between an effect on ATP association, ATP dissociation, or more complex kinetic effects. Nevertheless, the phosphorylation state of the R domain appears to regulate the interaction of the NBDs with ATP.

Our results show that the R domain does not function solely as an inhibitor that keeps the channel closed, so it is not simply an 'on-off' switch. Moreover, as individual phosphorylation sites do not contribute equally to activity, phosphorylation may have graded effects in stimulating the channel. Thus the function of the R domain differs mechanistically from that of the N-terminal 'ball' in the Shaker K^+ channel and the $ClC-2$ Cl^- channel, which are thought physically to obstruct the ion conduction pathway^{7,8,22}. How could the R domain exert both an inhibitory effect, which is removed by deletion, and a stimulatory effect, which is generated by phosphorylation? When phosphorylation modifies the R domain, it might have two effects: the first could be permissive, releasing a steric inhibition, consistent with our finding that deletion of the R domain produces a channel that is partially active; the second effect might be stimulatory, facilitating interaction of the NBDs with ATP, which would be consistent with results we obtained when a phosphorylated R domain was added to CFTR- $\Delta R/S660A$ and with the effects of CFTR-containing mutations in the endogenous R domain. By way of analogy, consider yeast glycogen phosphorylase, an enzyme that is activated when its amino terminus is phosphorylated^{23,24}. Deletion of the N terminus removes the phosphorylation site but generates an active enzyme, albeit one with much lower substrate affinity. Structural analysis showed that phosphorylation caused conformational changes which both removed a steric block from the catalytic site and stimulated an allosteric activation site.

Our data suggest a novel mechanism by which a protein domain can regulate ion channel activity; in CFTR, the phosphorylated R domain stimulates gating by a mechanism that is consistent with increased binding of ATP. We speculate that domains in other ion channels, and perhaps their modification by protein kinases, may stimulate the interaction with agonists rather than serving simply as inhibitory particles. □

Methods

Wild-type and mutant CFTR constructs were prepared and expressed in HeLa cells using a vaccinia virus hybrid expression system as described^{4,13,16}. For patch-clamp experiments, we used previously described methods¹³. The bath solution contained, in mM: 140 NMDG, 10 TES, 3 $MgCl_2$, 1 CsEGTA (pH adjusted to 7.3 with HCl); the pipette solution contained: 140 NMDG, 100 aspartic acid, 10 TES, 5 $CaCl_2$, 2 $MgCl_2$ (pH to 7.3 with HCl). All experiments were done at 37 °C with 1 mM ATP and 75 nM PKA on the cytosolic surface, except where noted. Channel activity was recorded and kinetics were evaluated as before^{13–15}. Closing times longer than 20 ms were used to define the end of a burst^{14,15}, although when times ranging from 15 to 40 ms were used it had very little effect on the results.

The R-domain peptide R1 (residues 645–834 of CFTR) was cloned into the vector pET21d which provides a C-terminal 6-His tag (Novagen), produced in *Escherichia coli* strain BL21pLysS, and purified by nickel-affinity chromatography according to manufacturer's protocol (pET His Tag System, Novagen). R1 was phosphorylated by addition of PKA and added to the bath solution at 50 nM. Kemptide and protein-kinase-inhibitor peptide from rabbit muscle were from Sigma.

Received 6 March; accepted 26 June 1997.

1. Welsh, M. J., Tsui, L.-C., Boat, T. F. & Beaudet, A. L. in *The Metabolic and Molecular Basis of Inherited*

- Disease* Vol. 7 (eds Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D.) 3799–3876 (McGraw-Hill, New York, 1995).
- Collins, F. S. Cystic fibrosis: molecular biology and therapeutic implications. *Science* **256**, 774–779 (1992).
 - Riordan, J. R. The cystic fibrosis transmembrane conductance regulator. *Annu. Rev. Physiol.* **55**, 609–630 (1993).
 - Rich, D. P. et al. Regulation of the cystic fibrosis transmembrane conductance regulator Cl^- channel by negative charge in the R domain. *J. Biol. Chem.* **268**, 20259–20267 (1993).
 - Ma, J. et al. Phosphorylation-dependent block of cystic fibrosis transmembrane conductance regulator chloride channel by exogenous R domain protein. *J. Biol. Chem.* **271**, 7351–7356 (1996).
 - Kartner, N. et al. Expression of the cystic fibrosis gene in non-epithelial invertebrate cells produces a regulated anion conductance. *Cell* **64**, 681–691 (1991).
 - Zagotta, W. N., Hoshi, T. & Aldrich, R. W. Restoration of inactivation in mutants of Shaker potassium channels by a peptide derived from ShB. *Science* **250**, 568–571 (1990).
 - Hoshi, T., Zagotta, W. N. & Aldrich, R. W. Biophysical and molecular mechanisms of Shaker potassium channel inactivation. *Science* **250**, 533–538 (1990).
 - Rich, D. P. et al. Effect of deleting the R domain on CFTR-generated chloride channels. *Science* **253**, 205–207 (1991).
 - Picciotto, M. R., Cohn, J. A., Bertuzzi, G., Greengard, P. & Nairn, A. C. Phosphorylation of the cystic fibrosis transmembrane conductance regulator. *J. Biol. Chem.* **267**, 12742–12752 (1992).
 - Anderson, M. P. et al. Nucleoside triphosphates are required to open the CFTR chloride channel. *Cell* **67**, 775–784 (1991).
 - Maller, J. E., Kemp, B. E. & Krebs, E. G. *In vivo* phosphorylation of a synthetic peptide substrate of cyclic AMP-dependent protein kinase. *Proc. Natl Acad. Sci. USA* **75**, 248–251 (1978).
 - Winter, M. C., Sheppard, D. N., Carson, M. R. & Welsh, M. J. Effect of ATP concentration on CFTR Cl^- channels: a kinetic analysis of channel regulation. *Biophys. J.* **66**, 1398–1403 (1994).
 - Carson, M. R., Travis, S. M. & Welsh, M. J. The two nucleotide-binding domains of cystic fibrosis transmembrane conductance regulator (CFTR) have distinct functions in controlling channel activity. *J. Biol. Chem.* **270**, 1711–1717 (1995).
 - Carson, M. R., Winter, M. C., Travis, S. M. & Welsh, M. J. Pyrophosphate stimulates wild-type and mutant CFTR Cl^- channels. *J. Biol. Chem.* **270**, 20466–20472 (1995).
 - Cheng, S. H. et al. Phosphorylation of the R domain by cAMP-dependent protein kinase regulates the CFTR chloride channel. *Cell* **66**, 1027–1036 (1991).
 - Cohn, J. A., Nairn, A. C., Marino, C. R., Melhus, O. & Kole, J. Characterization of the cystic fibrosis transmembrane conductance regulator in a colonocyte cell line. *Proc. Natl Acad. Sci. USA* **89**, 2340–2344 (1992).
 - Hwang, T. C., Nagel, G., Nairn, A. C. & Gadsby, D. C. Regulation of the gating of cystic fibrosis transmembrane conductance regulator Cl^- channels by phosphorylation and ATP hydrolysis. *Proc. Natl Acad. Sci. USA* **91**, 4698–4702 (1994).
 - Becq, F. et al. Phosphatase inhibitors activate normal and defective CFTR chloride channels. *Proc. Natl Acad. Sci. USA* **91**, 9160–9164 (1994).
 - Li, C. et al. ATPase activity of the cystic fibrosis transmembrane conductance regulator. *J. Biol. Chem.* **269**, 28463–28468 (1994).
 - Gunderson, K. L. & Kopito, R. R. Effects of pyrophosphate and nucleotide analogs suggest a role for ATP hydrolysis in cystic fibrosis transmembrane regulator channel gating. *J. Biol. Chem.* **269**, 19349–19353 (1994).
 - Gründer, S., Thiemann, A., Pusch, M. & Jentsch, T. J. Regions involved in the opening of $ClC-2$ chloride channel by voltage and cell volume. *Nature* **360**, 759–762 (1992).
 - Lin, K., Rath, V. L., Dai, S. C., Fletterick, R. J. & Hwang, P. K. A protein phosphorylation switch at the conserved allosteric site in GP. *Science* **273**, 1539–1541 (1996).
 - Lin, K., Hwang, P. K. & Fletterick, R. J. Mechanism of regulation in yeast glycogen phosphorylase. *J. Biol. Chem.* **270**, 26833–26839 (1995).

Acknowledgements. We thank P. Weber for preparing the cells, S. Travis for the R1 peptide, and G. Hill and D. Vermeer for assistance. This work was supported by the NHLBI and the HHMI. M.J.W. is an investigator of the HHMI.

Correspondence and requests for materials should be addressed to M.J.W. (e-mail: mjwelsh@blue.weeg.uiowa.edu).

Direct stimulation of Bruton's tyrosine kinase by G_q -protein α -subunit

Kendra Bence*, Wei Ma*, Tohru Kozasa† & Xin-Yun Huang

Department of Physiology, Cornell University Medical College, 1300 York Avenue, New York, New York 10021, USA

† Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas 75235, USA

* These authors contributed equally to this work.

Heterotrimeric guanine-nucleotide-binding regulatory proteins (G proteins) transduce signals from a wide variety of cell-surface receptors to generate physiological responses¹. Protein-tyrosine kinases are another group of critical cellular signal transducers and their malfunction often leads to cancer². Although activation of G-protein-coupled receptors can elicit rapid stimulation of cellular protein-tyrosine phosphorylation³, the mechanism used by G proteins to activate protein-tyrosine kinases is unclear. Here

we show that the purified α -subunit of the G_q class of G proteins ($G_{\alpha q}$) directly stimulates the activity of a purified non-receptor kinase, Bruton's tyrosine kinase (Btk)⁴, whereas purified α -subunits from G_{i1} , G_o or G_z proteins do not. $G_{\alpha q}$ can also activate Btk *in vivo*. Furthermore, in Btk-deficient cells, stimulation of another kinase, a p38 MAP kinase, by Gq-coupled receptors is blocked. Our results demonstrate that certain protein-tyrosine kinases can be direct effectors of G proteins.

In this reconstitution study using purified G-protein subunits and non-receptor tyrosine kinases, we show that $G_{\alpha q}$ can directly stimulate the non-receptor tyrosine kinase Btk. G-protein $G_{\alpha q}$, $G_{\alpha i1}$, $G_{\alpha o}$, $G_{\alpha z}$ and $G\beta_1\gamma_2$ subunits were purified from recombinant baculovirus-infected insect Sf9 cells⁵. The purity and identity of these G-protein subunits was confirmed by silver staining and western blotting of SDS-polyacrylamide gels (Fig. 1a, and data not shown). The activity of purified $G_{\alpha q}$ was assayed by its ability (at concentrations in the lower nanomolar range) to stimulate the phospholipase enzyme PLC- β 1 (data not shown). Recombinant human Btk was purified from *Escherichia coli* using an expression system that uses the bacterial chaperones GroES and GroEL to increase the yield of correctly folded, soluble, active proteins⁶; this

expression system has successfully produced large quantities of active recombinant human tyrosine kinases Csk and Lck^{6,7}. Purification from *E. coli*, which lacks endogenous protein-tyrosine kinases, ensures that Btk is correctly identified⁸. Btk was purified essentially free of other proteins, as judged by silver staining (Fig. 1a), and its identity was confirmed by western blotting with specific anti-Btk antibody (Fig. 1a).

Enzyme activity of purified Btk was assayed with a peptide substrate (amino-acid sequence: KKVVVALYDYMMPN) corresponding to residues 217–229 of human Btk, in which the tyrosine at residue 223 is the principal Btk autophosphorylation site⁹ (Fig. 1b). The activity of Btk is markedly stimulated when reconstituted with purified $G_{\alpha q}$ preactivated with GTP- γ S ($G_{\alpha q}$ -GTP γ S) (Fig. 1b). The time course of Btk-mediated substrate phosphorylation, in the absence (lanes 1–7) or presence (lanes 8–14) of $G_{\alpha q}$ -GTP γ S, is shown in Fig. 1b, c. Stimulation is dependent on the concentration of $G_{\alpha q}$ -GTP γ S (Fig. 1d). Because active GTP γ S-bound $G_{\alpha q}$ cannot be accurately measured, the stoichiometry of active $G_{\alpha q}$ versus Btk is not defined, but the amount of $G_{\alpha q}$ and Btk in our assay is similar to the amount of $G_{\alpha q}$ and PLC- β used previously^{10–12}. Three independent preparations of purified $G_{\alpha q}$ and several preparations

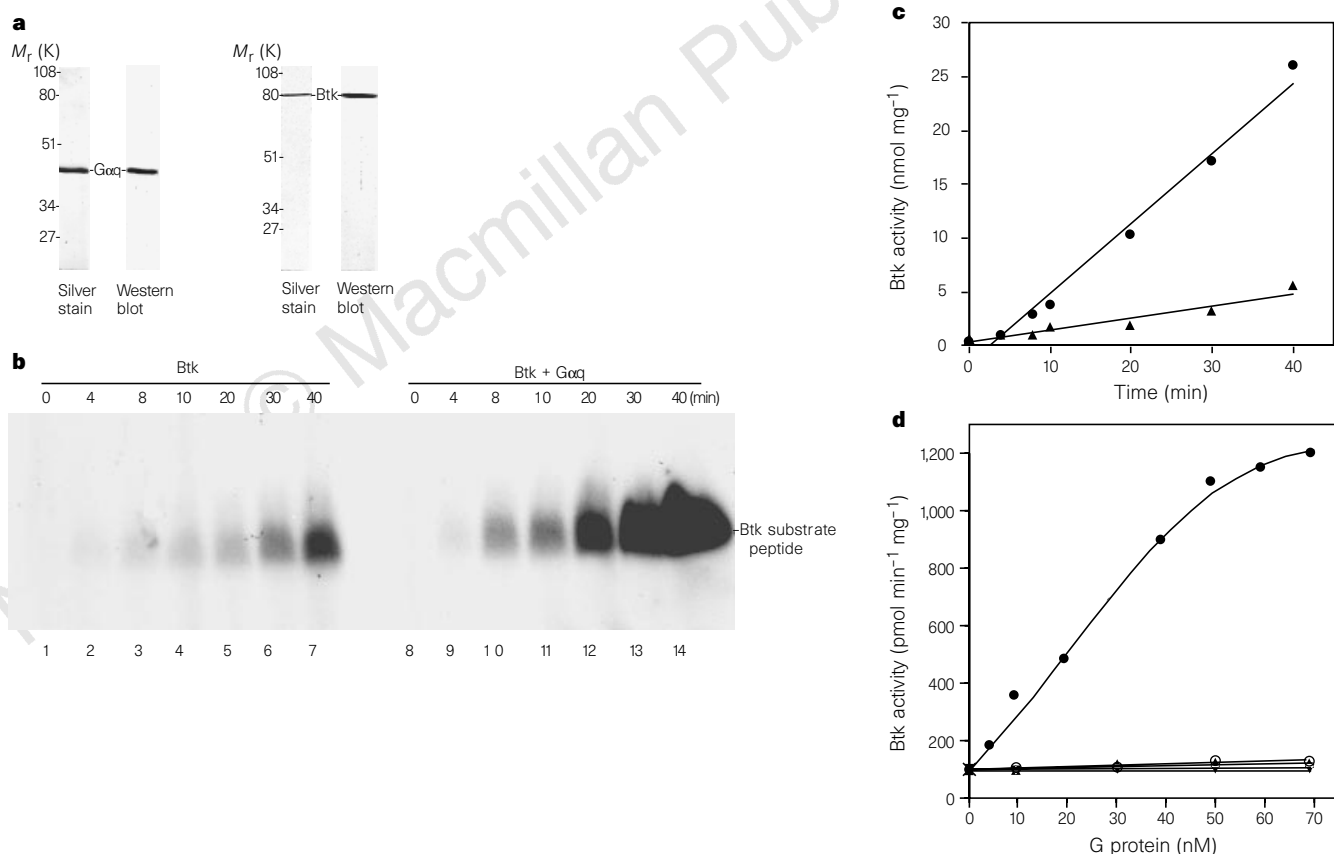


Figure 1 Stimulation of Btk kinase activity by $G_{\alpha q}$. **a**, Purification of $G_{\alpha q}$ and Btk. Left, silver staining of purified $G_{\alpha q}$ from recombinant baculovirus-infected Sf9 cells; western blotting confirmed the identity of $G_{\alpha q}$. Right, silver-staining of purified Btk from *E. coli*; the identity of Btk was confirmed by western blotting with anti-Btk antibody. Purified $G_{\alpha q}$ and Btk were subjected to SDS-PAGE. Gels were either stained with silver nitrate or the proteins were transferred to membrane filters for western blotting with specific antibodies. Prestained relative molecular mass (M_r) standards (BioRad) are shown on the left. **b** and **c**, Time course of Btk kinase activity. **b**, 1 ng Btk was incubated with 2 μ g substrate peptide for various times in the absence (lanes 1–7) or presence (lanes 8–14) of 20 nM $G_{\alpha q}$ -GTP γ S. After separation on 20% SDS-PAGE, signals were detected by autoradiography. **c**, After longer exposure, the corresponding bands on the gel were cut out and

phosphorylation quantified by scintillation counting. Btk activity was expressed as nmol PO_4 incorporated into peptide substrate per mg Btk. Data shown are representative of three similar experiments. **d**, Stimulation of Btk kinase by different concentrations of $G_{\alpha q}$ -GTP γ S. The indicated concentrations of α or $\beta\gamma$ subunits were reconstituted with 1 ng Btk for 20 min at 30°C. In the absence of $G_{\alpha q}$, Btk has a specific activity ~ 90 pmol $min^{-1} mg^{-1}$, close to that of Btk purified from mast cells²⁰. Increasing amounts of $G_{\alpha q}$ -GTP γ S (filled circles) increased Btk activity: for example, 50 nM $G_{\alpha q}$ -GTP γ S increases activity ~ 12 fold. $G_{\alpha i1}$ -GTP γ S (open triangle), $G_{\alpha q}$ -GDP (black triangles), $G\beta_1\gamma_2$ (open circles) or heat-inactivated $G_{\alpha q}$ -GTP γ S (black inverted triangle) had no effect on Btk activity. Data are representative of four similar experiments.

of purified Btk were tested and similar results were obtained with each. Reconstitution of Btk with purified $G\alpha i1$ -GTP γ S, $G\alpha o$ -GTP γ S, $G\alpha z$ -GTP γ S, $\beta_1\gamma_2$ or $G\alpha q$ -GDP did not alter Btk activity (Fig. 1d and Table 1). Moreover, heat inactivation of $G\alpha q$ -GTP γ S abolished its stimulatory effect on Btk (Fig. 1d and Table 1). Aluminium tetrafluoride (AlF_4) interacts with $G\alpha$ -bound GDP to mimic GTP and to activate G proteins¹³. AlF_4 -activated $G\alpha q$ -GDP increased Btk activity as expected, whereas AlF_4 alone had no stimulatory effect (Table 1), so we conclude that $G\alpha q$ directly stimulates Btk kinase activity.

Btk activity cannot be stimulated without $G\alpha q$: when $G\alpha q$ was immunoprecipitated out of a $G\alpha q$ -GTP γ S preparation with a $G\alpha q$ -specific antibody, the remainder could not stimulate Btk activity (Table 1). Depletion with antibodies against $G\alpha i$ and $G\alpha o$ did not abolish the stimulatory effect on Btk (data not shown).

To verify that Btk is responsible for phosphorylation, it was immunodepleted from the Btk preparation with a Btk-specific antibody and the remainder was tested for its ability to phosphorylate the peptide substrate, which it could not (Table 1).

We investigated whether $G\alpha q$ could activate Btk *in vivo* by studying the activation of endogenous Btk in lymphoma cells by $G\alpha q$. In DT40 lymphoma cells, Gq-coupled receptors, as represented by the well characterized Gq-coupled M1 muscarinic acetylcholine receptor (M1 mAChR), could activate endogenous Btk, as indicated by its increased kinase activity (Fig. 2a). A constitutively activated $G\alpha q$ mutant expressed in DT40 cells, $G\alpha qQ209L$, which

has a leucine residue at position 209 in the guanine-nucleotide-binding domain instead of glutamine, loses its GTPase activity and is thus a constitutively active form of $G\alpha q$ (ref. 14). Overexpression of $G\alpha qQ209L$ leads to increased kinase activity of endogenous Btk (Fig. 2b). $G\alpha q$ can therefore activate Btk *in vivo* as well as *in vitro*.

We investigated the physiological importance of $G\alpha q$ -mediated Btk activation in DT40 lymphoma cells by studying Gq signalling in Btk-deficient DT40 cells, which were generated by homologous targeting¹⁵. The kinase p38 is a member of a subclass of the mitogen-activated protein(MAP)-kinase family¹⁶ and can be activated by a mechanism involving a non-receptor-tyrosine-kinase¹⁷ and a G-protein-coupled receptor^{18,19}, so we tested whether Btk could regulate p38 signalling activated by Gq-coupled receptors in DT40 cells. As shown in Fig. 2c, Gq-coupled receptors can stimulate p38 kinase activity in wild-type DT40 cells, but not in Btk-deficient cells. To determine whether this failure was due to the absence of Btk, we transfected wild-type Btk back into Btk-deficient cells and found that p38 kinase could now be stimulated by Gq-coupled receptors (Fig. 2c). Our results indicate that Gq-mediated activation of Btk is necessary for stimulation of p38 kinase by Gq-coupled receptors in DT40 lymphoma cells. Hence, stimulation of Btk kinase by $G\alpha q$ may be essential *in vivo* for activation of p38 kinase.

How Btk is activated by $G\alpha q$ is not clear. The autophosphorylation site (Tyr 223) of Btk lies within its SH3 domain⁹. The SH3 domain of Itk (a member of the Btk kinase family) can interact with an adjacent proline-rich sequence to inhibit its association with other proteins²⁰, so $G\alpha q$ might promote a conformational change that disrupts this intramolecular interaction and so increase the autophosphorylation of Btk and hence its kinase activity towards its substrates. There is no sequence similarity between Btk and PLC- β , another downstream effector of $G\alpha q$ (refs 10, 11, 21, 22), which might otherwise be indicative of a possible $G\alpha q$ -binding site.

Btk is a member of a family of non-receptor tyrosine kinases^{4,23}, which include Btk, Itk/Tsk, Tec, Bmx and Txk²³ and are expressed in a variety of tissues. Defects in Btk are responsible for X-chromosome-linked agammaglobulinaemia in humans and X-chromosome-linked immunodeficiency in mice⁴. Given the high degree of sequence homology, it is likely that the other tyrosine kinases can also be directly stimulated by $G\alpha q$ and endogenous Gq-coupled receptors.

The pleckstrin-homology (PH) domain of Btk can bind $G\beta\gamma$ subunits, phospholipids, and several protein kinase C isoforms²⁴. We have previously shown that the $G\beta\gamma$ subunit, together with an unidentified membrane factor(s), possibly phospholipids, can increase the activity of immunoprecipitated Itk/Tsk (ref. 25); $G\beta\gamma$ alone barely affects the stimulation of Btk by $G\alpha q$ (Table 1), so the effect of $G\beta\gamma$ and phospholipids on Btk activity needs investigation.

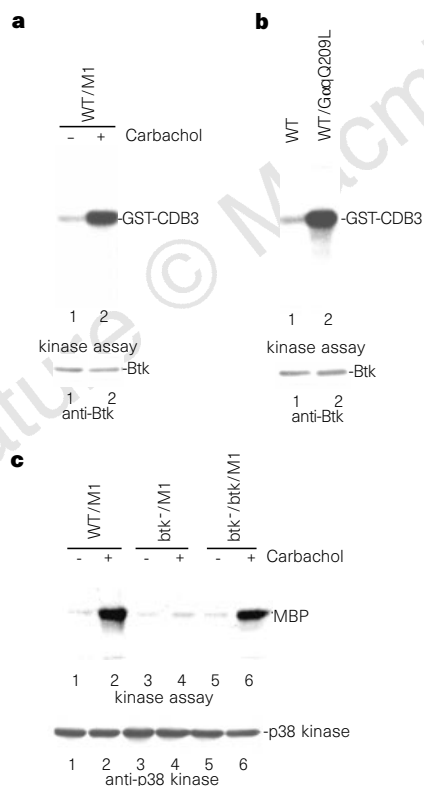


Figure 2 Stimulation of Btk kinase activity by Gq *in vivo*. **a**, Agonist (carbachol) activation of a Gq-coupled receptor (M1 mAChR) increases the kinase activity of endogenous Btk, as assayed with an exogenous substrate GST-CDB3. **b**, The ectopic expression of the constitutively activated $G\alpha q$ mutant $G\alpha qQ209L$ stimulates Btk kinase activity. **c**, Gq-coupled M1 mAChR activates p38 kinase in wild-type DT40 lymphoma cells but does not in Btk-deficient cells. Reconstitution of Btk into Btk-deficient cells restores stimulation of p38 kinase by Gq-coupled receptors. Myelin basic protein (MBP) was used as substrate for p38 kinase. **a-c**, Bottom, Western blots with anti-Btk or anti-p38 kinase antibodies showed that the amount of Btk or p38 kinase loaded in each lane was comparable.

Table 1 Effects on Btk kinase activity by various treatments

Treatment	Btk activity (pmol min ⁻¹ mg ⁻¹)
None	90
$G\alpha q$ -GTP γ S (50 nM)	1,100
Solvent	90
$G\alpha i1$ -GTP γ S (50 nM)	120
$G\alpha o$ -GTP γ S (50 nM)	100
$G\alpha z$ -GTP γ S (50 nM)	130
$\beta_1\gamma_2$ (100 nM)	120
$G\alpha q$ -GDP (50 nM)	130
Heat-inactivated $G\alpha q$ -GTP γ S (50 nM)	100
$G\alpha q$ - GTP γ S (50 nM) + $\beta_1\gamma_2$ (100 nM)	1,320
AlF_4 alone	70
AlF_4^- + $G\alpha q$ - GDP (50 nM)	1,200
Heat-inactivated $G\alpha q$ - GDP (50 nM) + AlF_4^-	90
Removal of Btk with antibody	0
Removal of $G\alpha q$ with antibody	100

None: Btk alone without G proteins. Solvent: the solution used to resuspend G-protein subunits⁵. Heat-inactivation: 90°C for 10 min. AlF_4^- : 10 mM NaF plus 30 μ M $AlCl_3$. Depletion with antibodies was done at 4°C for 4 h. Results are expressed as pmol phosphate incorporated into peptide substrate per min per mg of Btk. Data shown are from a representative experiment of three or four similar experiments.

We have shown that Gαq directly stimulates the kinase activity of Btk. To our knowledge, this is the first evidence indicating that a non-receptor tyrosine kinase can serve as a direct effector of heterotrimeric G proteins. As non-receptor tyrosine kinases are coupled to a variety of cellular signal transduction pathways, this G-protein-tyrosine-kinase link should provide a new means to extend G-protein signalling to a broad range of physiological processes. □

Methods

Protein purification. Purification of Gαq, Gαi1, Gαo, Gαz and Gβ₁γ₂ subunits from recombinant baculovirus-infected insect ovarian Sf9 cells has been described⁵. Briefly, the subunit was coexpressed with a His6-tagged associated subunit and after solubilization of membrane preparations, G-protein trimers were purified on a Ni²⁺-NTA column and the required subunit was dissociated with AIF₄; the subunit was further purified on an ion-exchange column. All antibodies against G-protein subunits were from Santa Cruz Biotechnology. Immunoprecipitation and western blotting have been described²⁶.

For purification of His6-tagged human Btk from *E. coli*, a human Btk cDNA was PCR-subcloned into vector pET21a with a His6 tag at its carboxyl end. The pET21a-H-Btk construct was then transformed into a bacterial strain BL21 that had been introduced with the bacterial chaperones GroES and GroEL (pREP4groESL)⁶. Eight litres of bacterial culture were grown at 37°C until the absorbance at 600 nm was ~0.7. Expression of Btk-His6 was then induced with IPTG (final concentration, 1 mM) (Research Products International) for 4 h at 37°C. The bacterial pellet was resuspended in lysis buffer (50 mM Tris, pH 7.4, 150 mM NaCl, 1% Triton X-100, 0.25 M sucrose, 1 mM EDTA, 1 mM EGTA, 1 mM sodium orthovanadate, 3 μM leupeptin, 3 μM pepstatin, 0.15 trypsin inhibitor units per ml aprotinin, 0.25 mM PMSF, 1 mM benzadimide) on ice. Lysozyme (1 mg ml⁻¹) was added and the sample incubated on ice for 30 min. After cell lysis by nitrogen cavitation (Parr bomb), the lysate was spun down at 10,000g for 20 min at 4°C. A slurry of Ni²⁺-NTA-agarose resin (4 ml, from Qiagen) was added to the supernatant after pre-equilibration of the resin with 50 mM Tris, pH 7.4, 300 mM NaCl. The mixture was gently agitated for 1 h at 4°C, then the beads were spun down and washed twice with 30 ml 50 mM Tris, pH 7.4, 300 mM NaCl, then poured into a column, washed with 50 mM Tris, pH 7.4, 250 mM NaCl, 5 mM imidazole until the absorbance at 280 nm was <0.01, then with 10 ml 50 mM Tris, pH 7.4, 250 mM NaCl, 25 mM imidazole, 15% glycerol. Proteins were eluted with 5 ml elution buffer (50 mM Tris, 300 mM NaCl, 250 mM imidazole, 10% glycerol, pH 6.0) and dialysed against Btk buffer (50 mM Tris, pH 7.4, 10 mM MnCl₂, 0.1% NP-40) and concentrated by using Millipore Ultrafree filters. Btk-His6 was further purified on a Mono-Q anion-exchange column for FPLC (Pharmacia Biotech Inc.) with a linear gradient of NaCl (0–400 mM; 40 ml) in Btk buffer plus 10% glycerol²⁷. Protein fractions were analysed by western blot and kinase assay. Anti-Btk monoclonal antibody was from PharMingen. Purity of Btk-His6 was verified by silver staining. Samples were concentrated, frozen in liquid nitrogen and stored at -80°C.

Kinase assays. Btk was assayed *in vitro* as follows: purified Btk kinase (1 ng) dialysed against kinase buffer (50 mM Tris, pH 7.4, 10 mM MnCl₂ and 10 μM ATP) was combined with 2 μg Btk peptide (KKVVALYDMPMN; Genemed) in the same buffer. The appropriate amount of purified G-protein subunit was added and kinase buffer used to bring the total reaction volume to 20 μl. After preincubation for 5 min at 30°C, 10 μCi [γ-³²P]ATP (3,000 Ci mmol⁻¹) was added and the mixture incubated at 30°C for 20 min or as indicated; the reaction was stopped by adding 20 μl 2× Laemmli sample buffer. The substrate peptide was separated on a 20% SDS-PAGE gel, dried and autoradiographed. Bands were cut out of the gel and counted in a scintillation counter.

Btk and p38 kinase assay in DT40 cells. DT40 avian B-lymphoma cells were grown as described^{15,26}. Transient transfections were carried out with 2 μg plasmid DNA and 5 μl Lipofectamine (Gibco-BRL) reagent in 6 well culture plates. Preparation of DT40 whole-cell extracts and immunoblotting were done as described^{15,26}. Expression of mAChRs in transfected cells was assessed by saturation ligand-binding analysis with intact cells using the muscarinic antagonist [³H]methylscopolamine²⁶. For p38 kinase assay, after 24 h transfected cells were starved in serum-free medium for 12 h and whole-cell extracts were made after stimulation with carbachol (100 μM) for 1 min²⁶. p38

immunoprecipitated with an antibody against p38 (Santa Cruz). p38 was assayed as described^{15,26}, using 10 μg MBP as substrate. The assay buffer was 30 mM Tris-HCl (pH 8), 20 mM MgCl₂, 2 mM MnCl₂, 10 μM ATP. The mixture was preincubated for 3 min before 10 μCi [γ-³²P]ATP was added. After 10 min at 30°C, samples were separated by 12% SDS-PAGE. Gels were transferred to nitrocellulose membrane filters and autoradiographed. For Btk assay, whole-cell extracts were prepared after stimulation with carbachol (100 μM for 20 s for m1 mAChR), Btk was immunoprecipitated by a monoclonal antibody (PharMingen), and Btk kinase activity was assayed with 5 μg of a fusion protein of purified glutathione-S-transferase (GST) with the cytoplasmic domain of human erythrocyte band-3 (GST-CDB3) as substrate. The GST-CDB3 plasmid was constructed by PCR cloning of CDB3 from the plasmid pCDB3/T7-7 into pGEX-2T (ref. 28). After 30 min at 30°C, samples were separated by 10% SDS-PAGE, gels were transferred to nitrocellulose and autoradiographed.

Received 2 June; accepted 20 June 1997.

1. Gilman, A. G proteins: transducers of receptor-generated signals. *Annu. Rev. Biochem.* **56**, 615–649 (1987).
2. Hunter, T. & Cooper, J. A. Protein-tyrosine kinases. *Annu. Rev. Biochem.* **54**, 897–930 (1985).
3. Hollenberg, M. D. Tyrosine kinase pathways and the regulation of smooth muscle contractility. *Trends Pharmacol. Sci.* **15**, 108–114 (1994).
4. Satterthwaite, A. & Witte, O. Genetic analysis of tyrosine kinase function in B cell development. *Annu. Rev. Immunol.* **14**, 131–154 (1996).
5. Kozasa, T. & Gilman, A. G. Purification of recombinant G proteins from Sf9 cells by hexahistidine tagging of associated subunits. *J. Biol. Chem.* **270**, 1734–1741 (1995).
6. Amrein, K. E., Takacs, B., Stieger, M., Molnos, J., Flint, N. A. & Burn, P. Purification and characterization of recombinant human p50csk protein-tyrosine kinase from an *Escherichia coli* expression system overproducing the bacterial chaperones GroES and GroEL. *Proc. Natl Acad. Sci. USA* **92**, 1048–1052 (1995).
7. Flint, N. A., Amrein, K. E., Jascu, T. & Burn, P. Purification and characterization of an activated form of the protein tyrosine kinase Lck from an *Escherichia coli* expression system. *J. Cell. Biochem.* **55**, 389–397 (1994).
8. Lindberg, R. A. & Pasquale, E. Isolation of cDNA clones that encode active protein-tyrosine kinases using antibodies against phosphotyrosine. *Meth. Enzymol.* **200**, 557–577 (1991).
9. Park, H. et al. Regulation of Btk function by a major autophosphorylation site within the SH3 domain. *Immunity* **4**, 515–525 (1996).
10. Smrcka, A. V., Hepler, J. R., Brown, K. O. & Sterweis, P. C. Regulation of polyphosphoinositide-specific phospholipase C activity by purified Gq. *Science* **251**, 804–807 (1991).
11. Lee, C.-W., Lee, K.-H., Lee, S.-B., Park, D. & Rhee, S. G. Regulation of phospholipase C-β4 by ribonucleotides and the α subunit of Gq. *J. Biol. Chem.* **269**, 25335–25338 (1994).
12. Hepler, J. R. et al. Functional importance of the amino terminus of Gαq. *J. Biol. Chem.* **271**, 496–504 (1996).
13. Bigay, J., Deterre, P., Pfister, C. & Chabre, M. Fluoride complexes of aluminium or beryllium act on G-proteins as reversibly bound analogues of the γ phosphate of GTP. *EMBO J.* **6**, 2907–2913 (1987).
14. Qian, N.-X., Winitz, S. & Johnson, G. L. Epitope-tagged Gα subunits: Expression of GTPase-deficient α subunits persistently stimulates phosphatidylinositol-specific phospholipase C but not mitogen-activated protein kinase activity regulated by the M1 muscarinic acetylcholine receptor. *Proc. Natl Acad. Sci. USA* **90**, 4077–4081 (1993).
15. Wan, Y. et al. Genetic evidence for a tyrosine kinase cascade preceding the mitogen-activated protein kinase cascade in vertebrate G protein signaling. *J. Biol. Chem.* **272**, 17209–17215 (1997).
16. Han, J., Lee, J. D., Bibbs, L. & Ulevitch, R. J. A MAP kinase targeted by endotoxin and hyperosmolarity in mammalian cells. *Science* **265**, 808–811 (1994).
17. Pandey, P. et al. Activation of p38 mitogen-activated protein kinase by c-Abl-dependent and -independent mechanisms. *J. Biol. Chem.* **271**, 23775–23779 (1996).
18. Kramer, R. M., Roberts, E. F., Striffler, B. A. & Johnstone, E. M. Thrombin induces activation of p38 MAP kinase in human platelets. *J. Biol. Chem.* **270**, 27395–27398 (1995).
19. Krump, E., Sanghera, J. S., Pelech, S. L., Furuya, W. & Grinstein, S. Chemotactic peptide N-formyl-Met-Leu-Phe activation of p38 mitogen-activated protein kinase (MAPK) and MAPK-activated protein kinase-2 in human neutrophils. *J. Biol. Chem.* **272**, 937–944 (1997).
20. Andreotti, A. H., Brunnell, S. C., Feng, S., Berg, L. J. & Schreiber, S. L. Regulatory intramolecular association in a tyrosine kinase of the Tec family. *Nature* **385**, 93–97 (1997).
21. Taylor, S. J., Chae, H. Z., Rhee, S. G. & Exton, J. H. Activation of the β1 isozyme of phospholipase C by α-subunits of the Gq class of G proteins. *Nature* **350**, 516–518 (1991).
22. Boyer, J. L., Graber, S. G., Waldo, G. L., Harden, T. K. & Garrison, J. S. Selective activation of phospholipase C by recombinant G-protein α and βγ subunits. *J. Biol. Chem.* **269**, 2814–2819 (1994).
23. Bolen, J. B. Protein tyrosine kinases in the initiation of antigen receptor signaling. *Curr. Opin. Immunol.* **7**, 306–311 (1995).
24. Lemmon, M. A., Ferguson, K. M. & Schlessinger, J. PH domains: diverse sequences with a common fold recruit signaling molecules to the cell surface. *Cell* **85**, 621–624 (1996).
25. Langhans-Rajasekaran, S. A., Wan, Y. & Huang, X.-Y. Activation of Tsk and Btk tyrosine kinases by G protein βγ subunits. *Proc. Natl Acad. Sci. USA* **92**, 8601–8605 (1995).
26. Wan, Y., Kurosaki, T. & Huang, X.-Y. Tyrosine kinases in activation of the MAP kinase cascade by G-protein-coupled receptors. *Nature* **380**, 541–544 (1996).
27. Price, D. J., Kawakami, Y., Kawakami, T. & Rivnay, B. Purification of a major tyrosine kinase from RBL-2H3 cells phosphorylating FcεRI γ-cytoplasmic domain and identification as the Btk tyrosine kinase. *Biochim. Biophys. Acta* **1265**, 133–142 (1995).
28. Wang, C. C. et al. Expression, purification, and characterization of the functional dimeric cytoplasmic domain of human erythrocyte band 3 in *E. coli*. *Protein Sci.* **1**, 1206–1214 (1992).

Acknowledgements. We thank M. Stieger for plasmid pREP4groESL; T. Kurosaki for Btk-deficient DT40 cells; P. Low for plasmid pCDB3/T7-7; and L. Levin, T. Maack, C. Malbon and members of our laboratory for reading the manuscript. Part of the Sf9 cells were cultured in the National Cell Culture Center which is supported by the National Center for Research Resources. This work was supported by grants from the NIH, the NSF and the American Heart Association. X.-Y.H. is a Cornell Scholar and a Beatrice F. Parvin Investigator of the American Heart Association New York City affiliate.

Correspondence and requests for materials should be addressed to X.-Y.H. (e-mail: xyhuang@med.cornell.edu).

A model for p53-induced apoptosis

Kornelia Polyak^{*†}, Yong Xia[‡], Jay L. Zweier[‡], Kenneth W. Kinzler[†] & Bert Vogelstein^{*†}

^{*} The Howard Hughes Medical Institute, [†] The Johns Hopkins Oncology Center, 424 North Bond Street, Baltimore, Maryland 21231, USA

[‡] Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

The inactivation of the *p53* gene in a large proportion of human cancers has inspired an intense search for the encoded protein's physiological and biological properties. Expression of *p53* induces either a stable growth arrest or programmed cell death (apoptosis). In human colorectal cancers, the growth arrest is dependent on the transcriptional induction of the protein *p21^{WAF1/CIP1}* (ref. 1), but the mechanisms underlying the development of *p53*-dependent apoptosis are largely unknown². As the most well documented biochemical property of *p53* is its ability to activate transcription of genes, we examined in detail the transcripts induced by *p53* expression before the onset of apoptosis. Of 7,202 transcripts identified, only 14 (0.19%) were found to be markedly increased in *p53*-expressing cells compared with control cells. Strikingly, many of these genes were predicted to encode proteins that could generate or respond to oxidative stress, including one that is implicated in apoptosis in plant meristems. These observations stimulated additional biochemical and pharmacological experiments suggesting that *p53* results in apoptosis through a three-step process: (1) the transcriptional induction of redox-related genes; (2) the formation of reactive oxygen species;

and (3) the oxidative degradation of mitochondrial components, culminating in cell death.

To evaluate the patterns of gene expression following *p53* expression, we used SAGE, a technique that allows the quantitative evaluation of cellular messenger RNA populations³. In brief, the method revolves around short sequence 'tags' (11 base pairs), generated from defined positions within each mRNA molecule. Expression patterns are deduced from the abundance of individual tags. To induce apoptosis, the colorectal cancer line DLD-1, containing an inactive endogenous *p53* gene, was infected with a replication-defective adenovirus encoding *p53* (Ad-*p53*). As previously shown, DLD-1 cells are among the ~50% of colorectal cancer (CRC) lines that undergo apoptosis in response to *p53* (ref. 4). RNA was purified from cells 16 h after infection, at least 8 h before the onset of morphological signs of apoptosis.

A total of 101,694 tags were analysed, approximately half from cells infected with Ad-*p53* and half from cells infected with a control virus (Ad-lacZ) encoding β -galactosidase. These tags corresponded to 7,202 different transcripts. Comparison of the two SAGE libraries indicated a remarkable similarity in expression profiles (Fig. 1a). Of the 7,202 transcripts detected, only 14 (0.19%) were expressed at levels more than 10-fold greater in *p53*-expressing than in control cells; conversely, only 20 transcripts were expressed at levels less than 10-fold lower in the *p53*-expressing cells (see <http://welchlink.welch.jhu.edu/~molgen-g/P53-SAGE.HTM>). As previous data indicated that *p53*-mediated transcriptional activation was the likely basis of *p53* action⁵, we concentrated on the 14 tags appearing at higher levels in the *p53*-expressing cells. The mRNA transcripts corresponding to 13 of these tags were successfully identified (Table 1), and in each case the induction was confirmed by northern blot analysis (examples shown in Fig. 1b). Only two of these genes (called *PIGs*, for *p53*-induced genes) had been implicated as targets of *p53*-transcriptional activation^{1,2,5,6}, and seven had not previously

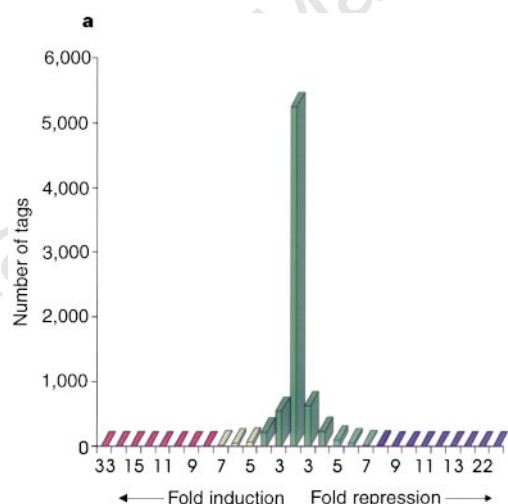
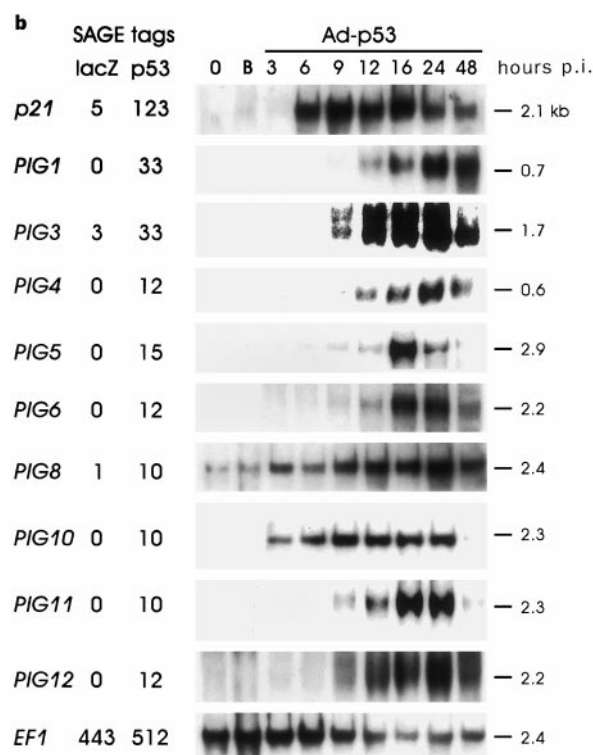


Figure 1 a, Summary of SAGE data. For each of 7,202 different transcripts identified, the ratio of their abundances in the two libraries is plotted. The y-axis indicates the number of tags expressed at the ratio indicated on the x-axis. Bars representing tags exhibiting less than 5-fold differences in expression are shown in green, and those induced or repressed more than 8-fold are shown in red and blue, respectively. **b**, Northern blot analysis after Ad-*p53* infection. Representative northern blots are shown for several transcripts identified by SAGE to be expressed at higher levels in *p53*-expressing cells at the indicated times post-infection (p.i.). Uninfected cells (column marked zero) and cells infected with Ad-lacZ for 48 h (column marked 'B') were included for comparison. EF1 is a control transcript expressed at relatively equal levels in cells 16 h after infection with Ad-*p53* and Ad-lacZ. The SAGE tag abundances (16 h after infection) are included on the left.



been described at all. Other genes previously implicated in p53-mediated responses were induced to lower levels (for example, *MDM2*, thrombospondin) or not at all (for example, *bax* and *cyclin G1*) in the human CRC cells studied here (ref. 4 and <http://welchlink.welch.jhu.edu/~molgen-g/P53-SAGE.HTM>).

PIG3 was induced at relatively short times after p53 expression, at least 12 h before any morphological or biochemical signs of apoptosis (Fig. 1b). This suggested that *PIG3* were directly induced by the transcriptional activation properties of p53. To test this conjecture formally in a representative case, we evaluated the genomic structure and sequence of *PIG3*. By screening a bacterial artificial chromosome (BAC) library, a genomic clone was identified that contained all *PIG3* coding sequences. The gene was localized to chromosome 2p (see Methods), and the intron-exon structure and sequence of the promoter region were determined. A 6.1-kilobase (kb) *Apa*I fragment of genomic DNA containing the presumptive promoter was then cloned upstream of a luciferase reporter gene (Fig. 2a). The resulting construct was transfected into three different human cell lines together with wild-type (WT) or mutant p53. As shown in Fig. 2b, wild-type p53 induced substantial activity through

the *PIG3* promoter in all three lines. Mutant p53 had no transcriptional activation capacity. Analysis of a truncated promoter showed that the p53-responsive elements lay within a fragment containing only 862 base pairs (bp) of sequence upstream of the *PIG3* transcription start site (Fig. 2a). Determination of the sequence of the 6.1-kb *Apa*I fragment (GenBank accession number AF010317) revealed a single 20-bp sequence predicted to bind p53, located at 308 nucleotides upstream of the transcription start site. A DNA fragment containing two copies of this sequence, but not a derivative of this fragment altered at critical residues, was found to bind strongly to p53 *in vitro* (Fig. 2c). As a further test of the p53-dependence of *PIG3* induction, we determined whether it could be induced by endogenous p53 rather than through the exogenous Ad-p53 source. Six CRC cancer cell lines were each treated with adriamycin, a DNA-damaging and apoptotic-inducing agent known to increase endogenous p53 levels. *PIG3*, like *p21*, was found to be strongly induced in the three lines with wild-type p53 genes, but not in the three lines with mutant p53 genes (<http://welchlink.welch.jhu.edu/~molgen-g/P53-SAGE.HTM>).

The sequences of the *PIG3* provided important clues as to their

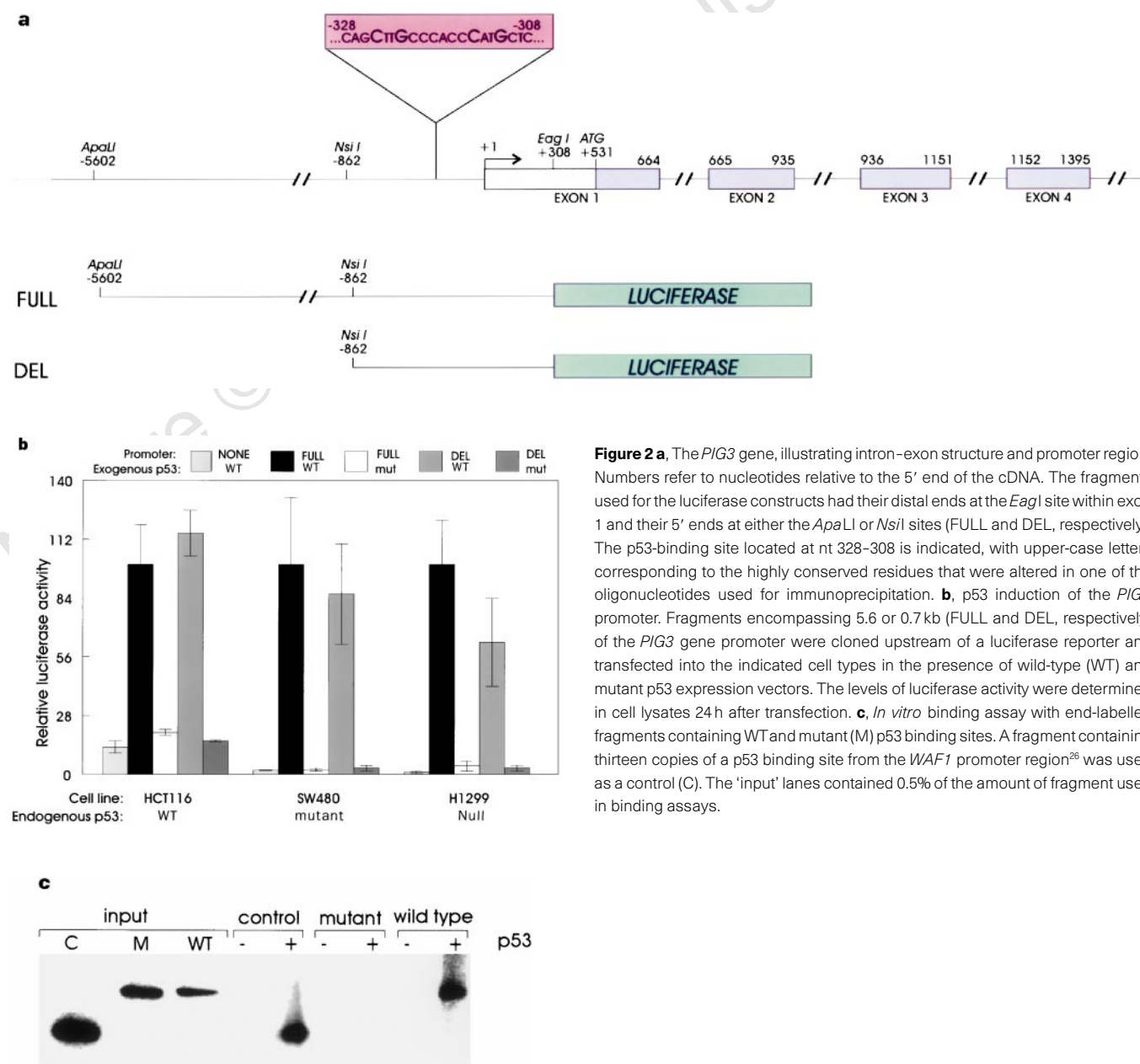


Figure 2 a, The *PIG3* gene, illustrating intron-exon structure and promoter region. Numbers refer to nucleotides relative to the 5' end of the cDNA. The fragments used for the luciferase constructs had their distal ends at the *Eag*I site within exon 1 and their 5' ends at either the *Apa*I or *Nsi*I sites (FULL and DEL, respectively). The p53-binding site located at nt 328-308 is indicated, with upper-case letters corresponding to the highly conserved residues that were altered in one of the oligonucleotides used for immunoprecipitation. **b**, p53 induction of the *PIG3* promoter. Fragments encompassing 5.6 or 0.7 kb (FULL and DEL, respectively) of the *PIG3* gene promoter were cloned upstream of a luciferase reporter and transfected into the indicated cell types in the presence of wild-type (WT) and mutant p53 expression vectors. The levels of luciferase activity were determined in cell lysates 24 h after transfection. **c**, *In vitro* binding assay with end-labelled fragments containing WT and mutant (M) p53 binding sites. A fragment containing thirteen copies of a p53 binding site from the *WAF1* promoter region²⁶ was used as a control (C). The 'input' lanes contained 0.5% of the amount of fragment used in binding assays.

potential functions (Table 1). In particular, several were predicted to encode proteins with activities related to the redox status of cells. *PIG12* is a novel member of the microsomal glutathione *S*-transferase family of genes (Fig. 3a). *PIG8* is the human homologue of a mouse gene (Ei24) whose expression is induced in a p53-dependent manner by etoposide, a quinone known to generate reactive oxygen species (ROS) (Fig. 3c). *PIG6* is a homologue of proline oxidoreductase (Fig. 3d), a mitochondrial enzyme that catalyses the first step in the conversion of proline to glutamate⁷. Glutamate is one of the three amino acids required for formation of glutathione, a major regulator of cellular redox status. The *p21* gene, which may also be considered to be a *PIG*, can be induced by reactive oxygen species, independently of p53 (ref. 8). *PIG4* encodes a serum amyloid protein that can be induced by oxidative stress⁹. *PIG1* belongs to the galectin family, members of which can stimulate superoxide production¹⁰. *PIG7* is induced by TNF- α , an inducer of oxidative stress. *PIG3* is a novel gene that is highly related to *TED2*, a plant NADPH oxidoreductase¹¹ (Fig. 3b). Interestingly, *TED2* is one of the few genes implicated in the apoptotic process necessary for the formation of plant meristems¹¹. The closest relative of *PIG3* in mammals is an NADPH-quinone oxidoreductase that is a potent generator of ROS¹².

Reactive oxygen species are powerful inducers of apoptosis¹³. The SAGE-based characterization of p53-induced genes suggested that p53 might induce apoptosis by stimulating the production of ROS. To test this hypothesis, the production of ROS was measured in p53-expressing cells using carboxymethyl dichlorofluorescein diacetate (CM-DCF-DA) and flow cytometry¹⁴. This analysis showed that

ROS were induced following Ad-p53 infection and that ROS continued to increase as apoptosis progressed (Fig. 4a). The magnitude of the increase in ROS, as assessed by CM-DCF-DA fluorescence, was comparable in p53-expressing cells and in cells treated with the powerful oxidant menadione (Fig. 4a). There was no change in CM-DCF-DA fluorescence following infection with a control adenovirus (Fig. 4a). As an assay for the functional consequences of ROS production, we examined the cellular content of cardiolipin, a major component of the mitochondrial membrane which is especially sensitive to cellular oxidation¹⁵. Using nonyl-acridine orange (NAO) as a probe, cardiolipin was found to decrease soon after p53-induced reactive oxygen species were detected (Fig. 4a), demonstrating significant injury to a major mitochondrial component.

To determine the specificity of *PIG* expression for the p53-dependent apoptotic process, we did experiments with other inducers of ROS or apoptosis. We found that *PIGs* were not expressed simply as a result of ROS production, as none were induced following treatment with menadione and only p21 was induced by hydrogen peroxide in DLD-1 cells (data not shown). Similarly, the specificity of *PIG* induction for p53-dependent apoptosis was confirmed by the demonstration that other inducers of apoptosis (indomethacin or ceramide) did not result in the expression of any *PIG*, despite extensive cell death (data not shown).

To clarify the relationship between p53 expression, *PIG* activation, ROS production, and apoptosis, we carried out more detailed time course experiments. *PIG* induction began within six hours of Ad-53 infection (Figs 1b and 4b), whereas intracellular ROS

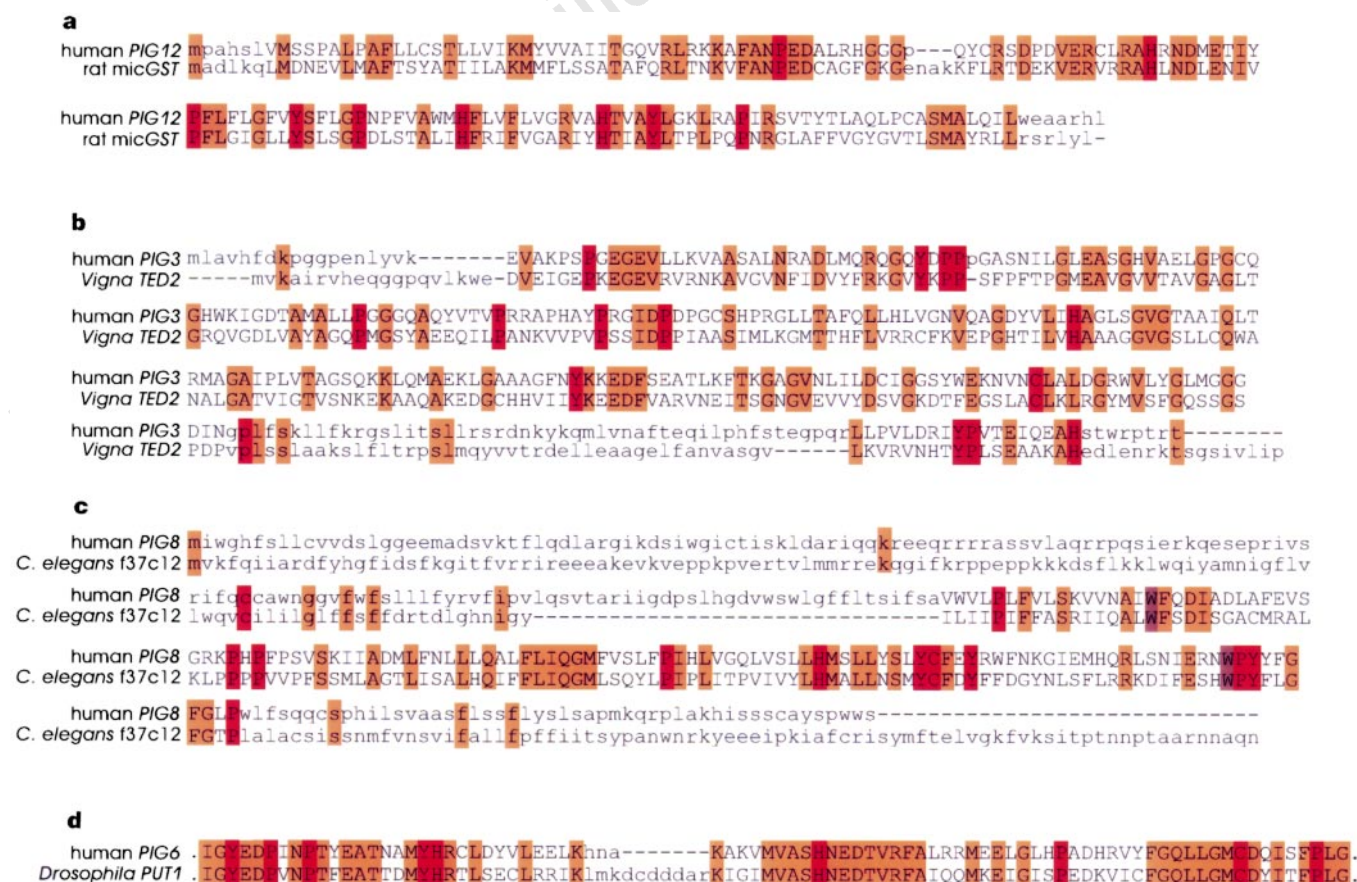


Figure 3 a-c, Sequences of selected genes identified through SAGE. In each case, the indicated gene is compared to the homologue from the non-human species that revealed a clue to its possible function. The amino-acid sequences were aligned using Macaw Version 2.0.3 and are shaded according to the mean

of their pairwise scores, with red indicating a higher score than orange. With the exception of *PIG6*, the cloned human sequences appeared to be full-length with respect to the coding region; accession numbers are given in Table 1.

production, as assessed with lucigenin chemiluminescence, could first be observed at 18 h (Fig. 4b). This ROS production led to oxidative stress, as evidenced by a $48 \pm 12\%$ decrease in intracellular glutathione concentration at 21 h. Mitochondrial lipid degradation (NAO staining) was not observed until three to six hours after the onset of a measurable ROS increase, and was accompanied by morphological (chromatin condensation and fragmentation) and biochemical (caspase-mediated degradation of polyADP-ribose polymerase (PARP)) signs of apoptosis (Fig. 4b). These observations are consistent with previous studies showing that mitochondrial damage is rapidly followed by classic signs of programmed cell death¹³.

The time courses shown in Fig. 4b suggest a cascade in which p53 transcriptionally induces redox-controlling genes, resulting in the production of ROS, in turn leading to oxidative damage to mitochondria and apoptosis. To determine whether these steps are causally associated, we inhibited each of them with specific pharmacological agents and determined the effect of this inhibition on other components of the pathway. First, cells were treated with the transcriptional inhibitor 5,6-dichlorobenimidazole riboside (DRB) at 8 h following Ad-p53 infection¹⁶. Although p53 expression was already near maximal at this time, DRB was found to block apoptosis at 24 h by $83 \pm 3\%$, as well as to inhibit the expression of *PIG3*. The translational inhibitor cycloheximide, when given up to 8 h following Ad-p53 infection, was found similarly to block apoptosis (by 79% at 24 h). Thus both transcription and translation were required for p53-induced apoptosis in CRC cells, as observed in some other systems and as expected for classic programmed cell death^{2,5}. Second, p53-expressing cells were treated with pyrrolidine

dithiocarbamate (PDTC), an antioxidant that blocks ROS-associated apoptosis¹⁷. PDTC was indeed able to block the apoptosis elicited by p53 (data not shown). However, PDTC inhibits many enzymes and its specificity is questionable¹⁷. We therefore treated cells with diphenyleneiodonium chloride (DPI), a specific inhibitor of flavin-dependent oxidoreductases that has been used to block production of ROS in a variety of systems¹⁸. Cells were treated with DPI 12 h after Ad-p53 infection, when *PIG3* production was already underway, and *PIG3* expression, apoptosis and ROS production were measured 12 h later. DPI (25 μ M) did not inhibit *PIG3* production, but inhibited ROS production by 71–85% and inhibited apoptosis by 73–77% in three independent experiments. Finally, we treated cells with bongkreikic acid (BA), a specific inhibitor of mitochondrial ATP translocase which can block the mitochondrial permeability transition pore opening thought to be required for some forms of apoptosis¹³. When cells were treated 12 h after Ad-p53 infection, BA did not inhibit either *PIG3* expression or ROS production, but inhibited subsequent apoptosis by 86–93%. BA was non-toxic at the dose used (100 μ M), and although BA inhibited the p53-apoptotic process dependent on ROS production, it had no effect on the p53-mediated growth arrest dependent on p21 (as assessed by flow cytometry).

Our gene-expression profile, time courses, and pharmacological inhibition studies strongly support a three-step model underlying p53's induction of apoptosis. We propose that p53 transcriptionally activates a specific subset of genes, including oxidoreductases, long before any morphological or biochemical evidence of cell death (Table 1 and Fig. 4b). The proteins encoded by these genes then collectively increase the content of ROS, which in turn damage

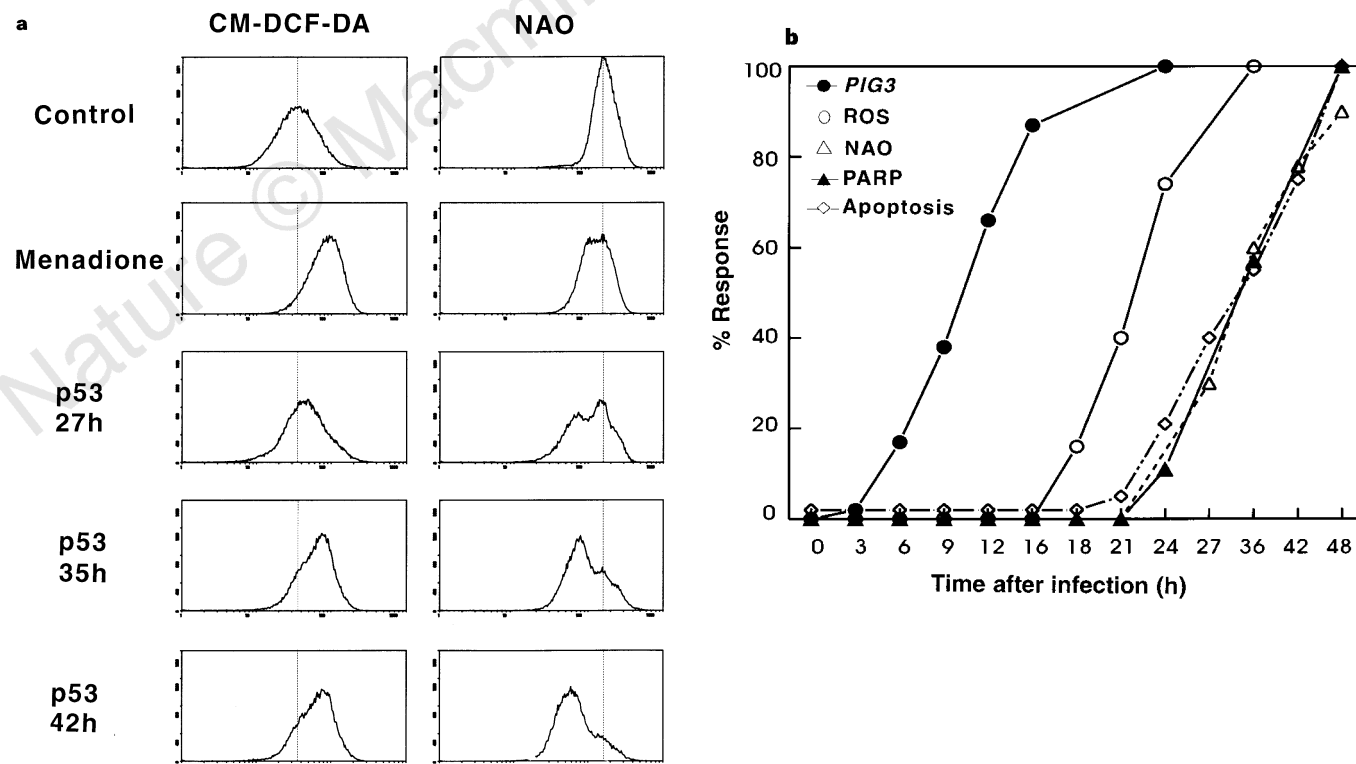


Figure 4 Oxidative stress and mitochondrial damage in p53-mediated apoptosis. **a**, DLD-1 cells were infected with Ad-p53 or control (Ad-lacZ) viruses and collected after 27, 35 or 42 h. Cells were incubated with CM-DCF-DA, a probe of ROS, or NAO, a probe of the mitochondrial membrane cardiolipin, and analysed by flow cytometry. The mean fluorescence of the control cells is indicated by vertical lines in each box. The pro-oxidant drug menadione was used as a positive control to induce oxidative stress. An increase in ROS and a decrease in cardiolipin concentration could be clearly observed by cytometry at 27 h and

increased as the p53-expressing cells entered apoptosis. **b**, Time course of apoptosis-related events following p53 expression. Cells were infected with Ad-p53 at 0 hours and *PIG3* expression (●) quantified by densitometry of northern blots. ROS production (○) was assessed with lucigenin; glutathione depletion exhibited a similar time course (not shown). Cardiolipin concentration (Δ) was assessed with NAO staining. Caspase activation (▲) was assessed by cleavage of PARP, and chromatin condensation/fragmentation (◊) was assessed by staining with DAPI.

Table 1 p53-induced genes (PIGs)*

Tag	Name	Function/homology	Accession no.
TGCTCTGGTTC	<i>p21</i>	CDK inhibitor	U03106
TAACCTCGTG	<i>PIG1</i>	Galectin-7	L07769
CCCTCTCCGT	<i>PIG2</i>	Guanidinoacetate <i>N</i> -methyl transferase	Z49878
GAGGCCAACA	<i>PIG3</i>	Quinone oxidoreductase homologue	AF010309
GTGCGGAGGAC	<i>PIG4</i>	Serum amyloid A	M10906
TGGGGCCGACG	<i>PIG5</i>	Normal keratinocyte mRNA	U33271
TCCTTGACCT	<i>PIG6</i>	Proline oxidase homologue	AF010310
CTGGGCCTGAA	<i>PIG7</i>	TNF- α induced mRNA	AF010312
AGCTGGTTTCC	<i>PIG8</i>	Etoposide-induced mRNA	AF010313
CGTCCCGGAGC	<i>PIG9</i>	Tax1-binding protein	U33822
AGCTCACTCCA	<i>PIG10</i>	Actin-binding protein	AF010314
AGGCTGTCCAG	<i>PIG11</i>	Unknown	AF010315
TGAGTCCCTGT	<i>PIG12</i>	Microsomal glutathione transferase homologue	AF010316
AGATGCTGCAG	<i>PIG13</i>	Unknown	NA

* Only those PIGs showing tenfold or greater reduction are presented. Additional p53-induced and repressed genes are available through <http://welchlink.welch.jhu.edu/~molgen-g/p53-SAGE.HTM>.

mitochondria. Leakage of calcium and proteinaceous components from damaged mitochondria then stimulate the caspases that are ubiquitously activated during the apoptotic process^{19–22}. Data from several experimental systems are consistent with this model. For example, apoptosis induced by irradiation, which is dependent on p53 in certain cell types, has been proposed to proceed through a process involving ROS and mitochondrial damage²³. Additionally, an SV40 large-T-antigen mutant, which binds p53 only at the permissive temperature, induces apoptosis at the non-permissive temperature through a ROS-related mechanism²⁴. Also, p53-induced apoptosis in smooth-muscle cells is ROS-dependent²⁵. Thus, the events we observed in CRC cells are unlikely to be cell-type- or species-specific, and may often underlie p53-associated apoptotic processes. The fact that one of the PIGs is highly related to *Ted2*, an oxidoreductase implicated in plant cell apoptosis¹¹, and that apoptosis in plants may also proceed through a ROS-directed pathway¹¹, adds further interest to this model.

Although observations by us and others are consistent with this model, they raise several unanswered questions. For example, we do not yet know which of the PIGs are primarily responsible for the induction of ROS. We suspect that their combination, rather than any single one, is necessary for ROS generation. This conjecture is supported by preliminary experiments demonstrating that *PIG3* alone does not induce apoptosis when overexpressed (K.P., unpublished data). Although we have concentrated on the most highly induced PIGs, the SAGE analysis revealed at least 26 other genes that were induced by p53 to significant but to lower extents than *p21* and *PIG1–PIG13*. Some of these 26 genes (available through <http://welchlink.welch.jhu.edu/~molgen-g/p53-SAGE.HTM>) may play a role in redox regulation. It is also not known why some cells enter into apoptosis following p53 expression while others undergo a prolonged growth arrest⁴. The possibility that PIGs are only induced in the former has been excluded by examination of PIG expression in such lines; most PIGs were induced by p53 in each of ten CRC lines tested, regardless of whether they underwent apoptosis or growth arrest. A more likely possibility is that different cells have different capacities to cope with generators of oxidative stress, and that cells with a low capacity succumb to apoptosis. This possibility is supported by studies that show that the response to ROS varies significantly with cell type and growth conditions¹³. We hope that the experiments and genes described here will open a new window into the p53 apoptotic process and will facilitate enquiry into these issues. □

Methods

Cells and RNA. All cell lines were obtained from the American Type Culture Collection and were cultured in McCoy's medium supplemented with 10% fetal bovine serum (FBS). Cells were infected with recombinant adenoviruses containing either the p53 gene or the β -galactosidase gene²⁶ at a multiplicity of infection of 10–100. RNA was purified from cells at various times after

infection, using the MessageMaker Kit (Gibco/BRL). Northern blot analysis has been described²⁶.

SAGE. SAGE was done as before^{3,27}. In brief, polyadenylated RNA was converted to double-stranded cDNA with a BRL synthesis kit following the manufacturer's protocol, with the inclusion of primer biotin-5'-T₁₈-3'. The cDNA was cleaved with *Nla*III and the 3'-terminal cDNA fragments were bound to streptavidin-coated magnetic beads (Dynal). After ligation of oligonucleotides containing recognition sites for *Bsm*FI, the linked cDNA was released from the beads by digestion with *Bsm*FI. The released tags were ligated to one another, concatemerized, and cloned into the *Sph*I site of pZero 1.0 (Invitrogen). Colonies were screened with PCR using M13 forward and M13 reverse primers. PCR products containing inserts of greater than 300 bp (>20 tags) were sequenced with the TaqFS DyePrimer kit and analysed using a 377 ABI automated sequencer (Perkin Elmer).

Statistical analysis. 53,022 and 51,853 tags were identified from DLD-1 cells infected with Ad-p53 and Ad-lacZ, respectively. The two libraries were compared using the SAGE program group³. Corrections for tags containing linker sequences and other potential artefacts were made as described²⁷. Of 104,875 total tags identified, 3,181 were excluded from analysis on this basis. Monte Carlo simulations revealed that the computational analyses had a >99% probability of detecting a transcript expressed at an abundance of 0.00005 in either RNA sample.

cDNA clones. Cellular mRNA from Ad-p53-infected cells was used to prepare cDNA as described for the SAGE libraries except that the 3' primer contained an additional M13 forward sequence between the oligo(dT) tract and the biotinylated 5' residue. To determine the sequence of the transcript from which an individual tag was derived, this cDNA was used as a template for PCR, employing an M13 forward primer and a primer containing the tag sequence. In other cases, mRNA from Ad-p53-infected cells was used to construct a cDNA library in the ZAP Express vector (Stratagene) and the library was screened by hybridization with oligonucleotides corresponding to tags, as described³. Of 14 tags identified by SAGE as differentially expressed in p53-expressing cells, 8 corresponding genes could be identified simply by searching public databases, particularly those including expressed sequence tags. In 5 cases, one of the two strategies described was used to obtain the corresponding PIG. In one of the 14 cases (*PIG13*), no cDNA clone could be recovered corresponding to the tag sequence.

Analysis of *PIG3* genomic structure. An arrayed BAC library (Research Genetics) was screened by PCR using the following primers derived from the 5' end of the *PIG3* gene: 5'-ggc-cag-gag-taa-gta-act-3' and 5'-gcc-ctg-gtc-tgc-cgc-gga-3'. *Eco*RI fragments encompassing the *PIG3* coding sequences were subcloned into pBR322 and partially sequenced to determine the intron–exon borders. A 6.1-kb *Apa*LI fragment whose 3' end was at a *Eag*I site 308 bp downstream of the transcription start site was then cloned into a promoterless luciferase reporter vector (Fig. 2a). This fragment was completely sequenced by primer walking. Subclones were then generated by restriction endonuclease digestion. Luciferase activity was determined after co-transfection with expression vectors encoding wild-type or R175H mutant p53. For *in vitro* p53 binding experiments, oligonucleotides containing two copies of the predicted p53-binding site (Fig. 2a) were subcloned into a modified pBR322 vector, excised as

a ~260-bp restriction fragment, and end-labelled. Immunochemical assays were performed as described²⁸.

Flow cytometry and other assays. Cells were collected with the aid of trypsin and incubated with CM-DCF-DA or NAO (Molecular Probes) at concentrations of 10 and 0.4 μ M, respectively, for 20 min at 37°C before analysis by flow cytometry^{14,15}. To determine the fraction of apoptotic cells after various treatments, cells were stained with the DNA-binding dye H33258 and evaluated by fluorescence microscopy or flow cytometry as described¹. Superoxide production was assessed with lucigenin²⁹. In brief, $4\text{--}5 \times 10^6$ cells were collected with a rubber policeman and resuspended in 1 ml Earle's balanced salt solution (Life Technologies). Dark-adapted lucigenin (bis-*N*-methylacridinium nitrate; Sigma) was added to the samples to 20 μ M final concentration. Light emission was detected using a Berthold LB 9505 luminometer for 60 min at 37 °C. Glutathione concentrations were measured using an assay kit (Oxford Biomedical) according to the manufacturer's instructions. Caspase activation was assessed by cleavage of PARP. Lysates from cells infected with Ad-p53 were western-blotted with an anti-PARP antibody and the cleavage fragments quantified by densitometry⁴.

Received 27 May; accepted 1 July 1997.

- Waldman, T., Kinzler, K. W. & Vogelstein, B. p21 is necessary for the p53-mediated G1 arrest in human cancer cells. *Cancer Res.* **55**, 5187–5190 (1995).
- Oren, M. Relationship of p53 to the control of apoptotic cell death. *Semin. Cancer Biol.* **5**, 221–227 (1994).
- Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
- Polyak, K., Walkman, T., He, T.-C., Kinzler, K. W. & Vogelstein, B. Genetic determinants of p53 induced apoptosis and growth arrest. *Genes Dev.* **10**, 1945–1952 (1996).
- Levine, A. J. p53, the cellular gatekeeper for growth and division. *Cell* **88**, 323–331 (1997).
- Lehar, S. M. *et al.* Identification and cloning of E124, a gene induced by p53 in etoposide-treated cells. *Oncogene* **12**, 1181–1187 (1996).
- Hayward, D. C. *et al.* The *sluggish-A* gene of *Drosophila melanogaster* is expressed in the nervous system and encodes proline oxidase, a mitochondrial enzyme involved in glutamate biosynthesis. *Proc. Natl Acad. Sci. USA* **90**, 2979–2983 (1993).
- Russo, T. *et al.* A p53-independent pathway for activation of WAF1/CIP1 expression following oxidative stress. *J. Biol. Chem.* **270**, 29386–29391 (1995).
- Reinhoff, H. Y. Jr, Huang, J. H., Li, X. X. & Liao, W. S. Molecular and cellular biology of serum amyloid A. *Mol. Biol. Med.* **7**, 287–298 (1990).
- Yamaoka, A., Kuwabara, I., Frigeri, I. G. & Liu, F. T. A human lectin, galectin-3 (epsilon bp/Mac-2) stimulates superoxide production by neutrophils. *J. Immunol.* **154**, 3479–3487 (1995).
- Greenberg, J. T. Programmed cell death: A way of life for plants. *Proc. Natl Acad. Sci. USA* **93**, 12094–12097 (1996).
- Rao, P. V., Krishna, C. M. & Zigler, J. S. Jr Identification and characterization of the enzymatic activity of zeta-crystallin from guinea PIG lens. A novel NADPH: quinone oxidoreductase. *J. Biol. Chem.* **267**, 96–102 (1992).
- Kroemer, G., Zamzami, N. & Susin, S. A. Mitochondrial control of apoptosis. *Immun. Today* **18**, 45–51 (1997).
- Zamzami, N. *et al.* Reduction in mitochondrial potential constitutes an early irreversible step of programmed lymphocyte death *in vivo*. *J. Exp. Med.* **181**, 1661–1672 (1995).
- Petit, P. X. *et al.* Alterations in mitochondrial structure and function are early events of dexamethasone-induced thymocyte apoptosis. *J. Cell. Biol.* **130**, 157–167 (1995).
- Tamm, I. & Sehgal, P. B. Halobenzimidazole ribosides and RNA synthesis of cells and viruses. *Adv. Virus Res.* **22**, 187–258 (1978).
- Orrenius, S., Nobel, C. S. I., van den Dobbelen, D. J., Burkitt, M. J. & Slater, A. F. G. Dithiocarbamates and the redox regulation of cell death. *Biochem. Soc. Transact.* **24**, 1032–1038 (1996).
- Holland, P. C., Clark, M. G., Bloxham, D. P. & Lardy, H. A. Mechanism of action of the hypoglycemic agent diphenyliodonium. *J. Biol. Chem.* **248**, 6050–6056 (1973).
- Korsmeyer, S. J. Regulators of cell death. *Trends Genet.* **11**, 101–105 (1995).
- Susin, S. A. *et al.* Bcl-2 inhibits the mitochondrial release of an apoptogenic protease. *J. Exp. Med.* **184**, 1331–1341 (1996).
- Yang, J. *et al.* Prevention of apoptosis by Bcl-2: release of cytochrome c from mitochondria blocked. *Science* **275**, 1129–1132 (1997).
- Kluck, R. M., Bossy-Wetzell, E., Green, D. R. & Newmeyer, D. D. The release of cytochrome c from mitochondria: a primary site for Bcl-2 regulation of apoptosis. *Science* **275**, 1132–1136 (1997).
- Borek, C. Radiation and chemically induced transformation: free radicals, antioxidants and cancer. *Br. J. Cancer suppl.* **8**, 74–86 (1987).
- Vayssiere, J. L., Petit, P. X., Risler, Y. & Mignotte, B. Commitment to apoptosis is associated with changes in mitochondrial biogenesis and activity in cell lines conditionally immortalized with simian virus 40. *Proc. Natl Acad. Sci. USA* **91**, 11752–11756 (1994).
- Johnson, T. M., Yu, Z.-X., Ferrans, V. J., Lowenstein, R. A. & Finkel, T. Reactive oxygen species are downstream mediators of p53-dependent apoptosis. *Proc. Natl Acad. Sci. USA* **93**, 11848–11852 (1996).
- El-Deiry, W. S. *et al.* WAF1, a potential mediator of p53 tumor suppression. *Cell* **75**, 817–825 (1993).
- Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
- El-Deiry, W. S., Kern, S. E., Pietenpol, J. A., Kinzler, K. W. & Vogelstein, B. Definition of a consensus binding site for p53. *Nature Genet.* **1**, 45–49 (1992).
- Faulkner, K. & Fridovich, I. Luminol and lucigenin as detectors for O₂⁻. *Free Rad. Biol. Med.* **15**, 447–451 (1993).

Acknowledgements. We thank V. Velculescu, L. Zhang and W. Zhou for help with SAGE analysis; J. A. Duine for bongkrekic acid; J. Flook for assistance with flow cytometry; and members of our laboratories for discussion. This work was supported by the Clayton Fund and by grants from the NIH. K.P. is a research associate and B.V. is an investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to K.W.K.

Self-organization of microtubules and motors

F. J. Nédélec^{*†}, T. Surrey^{*}, A. C. Maggs[†] & S. Leibler^{*}

^{*} Departments of Molecular Biology and Physics, Princeton University, Princeton, New Jersey 08544, USA

[†] Laboratoire de Physico-Chimie Théorique, Ecole Supérieure de Physique et Chimie Industrielles, 10 rue Vauquelin, 75231 Paris, France

Cellular structures are established and maintained through a dynamic interplay between assembly and regulatory processes. Self-organization of molecular components provides a variety of possible spatial structures: the regulatory machinery chooses the most appropriate to express a given cellular function¹. Here we study the extent and the characteristics of self-organization using microtubules and molecular motors² as a model system. These components are known to participate in the formation of many cellular structures, such as the dynamic asters found in mitotic and meiotic spindles^{3,4}. Purified motors and microtubules have previously been observed to form asters *in vitro*⁵. We have reproduced this result with a simple system consisting solely of multi-headed constructs of the motor protein kinesin⁶ and stabilized microtubules. We show that dynamic asters can also be obtained from a homogeneous solution of tubulin and motors. By varying the relative concentrations of the components, we obtain a variety of self-organized structures. Further, by studying this process in a constrained geometry of micro-fabricated glass chambers⁷, we demonstrate that the same final structure can be reached through different assembly 'pathways'.

A dividing cell has to separate spatially its newly duplicated chromosomes. It uses the bipolar spindle, a complex molecular structure made of many different protein components, to accomplish this task with high precision. Spindle formation is an intensively studied prototype of cell morphogenesis phenomena. A bipolar spindle forms when mitotic chromosomes capture and selectively stabilize dynamic microtubules nucleated by two centrosomes¹. This assembly process tolerates numerous variations in intermediate configurations, such as different chromosome positions and geometrical and mechanical perturbations⁸. Spindles

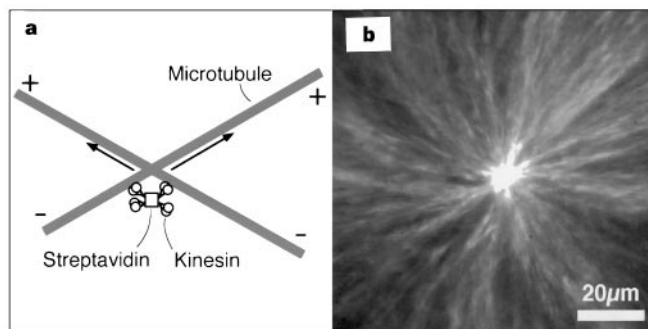


Figure 1 Self-organization of taxol-stabilized microtubules and kinesin constructs into asters. **a**, Schematic representation of a kinesin-streptavidin construct moving simultaneously along two microtubules. The kinesin constructs can be seen as force-generating, mobile crosslinks. **b**, A self-organized aster observed by dark-field microscopy. Polymerized and taxol-stabilized microtubules were mixed with motor constructs and ATP. The sample is shown after ~2 min. The bright spot in the centre of the aster is caused by light scattering from accumulated microtubules and motors. In cells and cell-free extracts, where the asters are organized by minus-end directed motors such as dyneins^{16,17}, the aster polarity is opposite. Scale bar, 20 μ m.

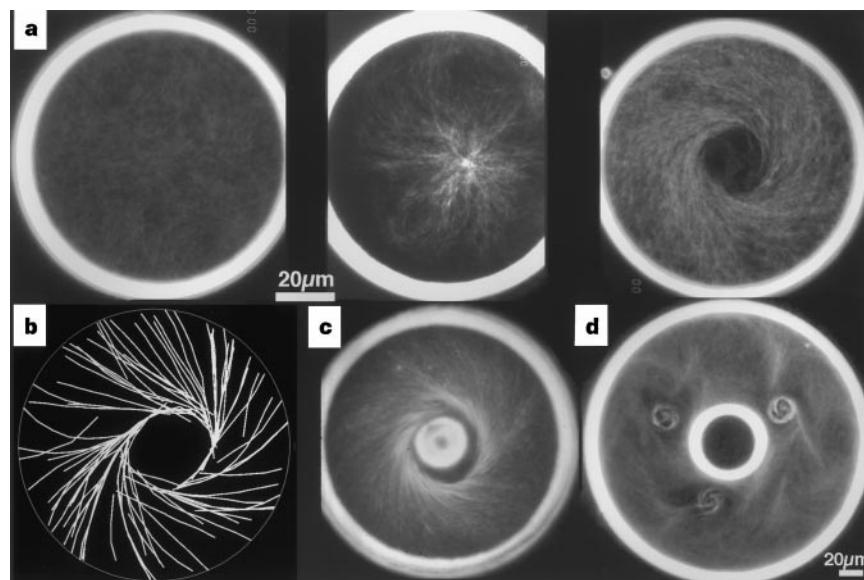


Figure 2 Self-organization in the constrained geometry of micro-fabricated chambers etched in glass. **a**, Formation of a vortex as observed by dark-field microscopy in a chamber of diameter $90\mu\text{m}$ and depth $5\mu\text{m}$. Microtubule polymerization has been initiated by warming to 37°C . Left, after ~ 0.5 min, microtubules nucleate uniformly in the sample; middle, after ~ 1.5 min, an aster forms in the centre of the chamber; right, after ~ 3 min, a steady-state vortex structure is observed. **b**, Two-dimensional numerical simulation of the same process. The microtubule mean length was initially smaller than the chamber radius, but became longer than it in the state shown. **c**, Vortex structure in a torus-shaped chamber of the same diameter as previously, but including a central unetched region. **d**, Steady-state structure in a torus-shaped chamber of larger size ($240\mu\text{m}$). The three vortices have the same core size as those in **a** and **c**. Scale bars, $20\mu\text{m}$.

also form in the absence of centrosomes, as observed in meiotic cell divisions, and, more surprisingly, in *in vitro* experiments in *Xenopus laevis* egg extracts devoid of centrosomes^{9,10}, in which case microtubules grown in the vicinity of chromatin are organized into bipolar spindle by molecular motors.

In order to explore the issues of self-organization of microtubules and motors, we have constructed an extremely simplified system with only a few purified components. In addition to tubulin, this system includes artificial molecular constructs of several kinesin heads associated through biotin–streptavidin links. We chose kinesin because of its extensive molecular characterization^{11–13}. The motor constructs (kinesins or motors) attach to neighbouring microtubules and, in the presence of ATP, move towards the microtubule ‘plus’ ends¹⁴. In this way they form dynamic crosslinks between microtubules (Fig. 1a).

An aster formed in a system of microtubules stabilized with taxol¹⁵ and mixed with motors is shown in Fig. 1b. The sample is sandwiched between slide and coverslip in a quasi-two-dimensional geometry. The starting configuration is an isotropic ‘gel’ consisting of microtubules crosslinked randomly by motors (not shown). Asters of microtubules form within a few minutes as kinesins accumulate in their centres. The size of the asters is expected to be determined by the length distribution of stabilized microtubules. Similar asters were previously observed to form in a system of purified heavy chains of kinesin mixed with microtubules⁵, as well as in *Xenopus laevis* egg extracts with the addition of microtubule stabilizing agents^{16,17}.

The use of taxol allows us to control microtubule length and number, although stabilizing microtubules with taxol is not necessary for aster formation: asters also form if microtubules are allowed to assemble and disassemble through the dynamic instability¹⁸. All further experiments, presented below, are performed in the absence of taxol.

To study the influence of geometrical constraints on self-organization, we have encapsulated solutions of tubulin and motors in sealed micro-fabricated glass chambers⁷ of various shapes, with typical lateral dimensions of $100\mu\text{m}$ and a thickness of $5\mu\text{m}$ (Fig. 2). In a cylindrical geometry, microtubules polymerizing from an initially homogeneous solution first organize into a symmetric aster centred in the chamber. As microtubules continue to grow and begin to buckle, the centre of the aster becomes unstable and a vortex structure forms. Vortices can be observed simultaneously in hundreds of chambers etched in a single coverslip and filled with the same initial solution. We saw an equal number of

structures with right-handed and left-handed vorticity. Because vortices form from asters, the microtubules have their plus ends oriented towards the centre of the vortex. Small objects, such as short microtubules or small colloidal beads with motors attached to them, circle around the core of the vortex. This circulatory motion is reminiscent of the microtubule- and motor-dependent movement of cytoplasm observed in the development of *Drosophila melanogaster* oocytes¹⁹.

One advantage of using highly simplified systems with a small number of well-characterized components is the accessibility to quantitative theoretical analysis. Aster formation can be studied in numerical simulations, in which microtubules are treated as flexible, polar rods, and kinesin-like motors are characterized by a linear force–velocity curve^{11,12} and high processivity^{11,12}. Vortices can also be reproduced in numerical simulation by incorporating geometrical confinement and the dynamic assembly of microtubules (Fig. 2b). This simulation exhibits an assembly pathway very similar to that in the experiments, in which an initially symmetric aster is destabilized by the growth and buckling of microtubules.

To determine whether assembly through a symmetric aster is the only possible route leading to the final vortex structure, we enclosed the same solution of proteins in a chamber of the same size but with the topology of a torus. The idea for this came from experiments of microsurgery performed on melanophores²⁰ in which the geometry of the cell was changed artificially. The formation of symmetric asters is now prevented by the geometry of the chamber, but the same final steady state, a vortex, can nevertheless be reached (Fig. 2c), demonstrating that this simple system can find alternative ‘assembly pathways’. This is similar to results obtained in *Xenopus laevis* egg extracts, in which microtubule asters were formed either around centrosomes or through motor-mediated assembly^{9,10}. The existence of alternative ‘assembly pathways’ can be of more general importance. It is easy to imagine that, as in the case of metabolic or signal transduction pathways, a given route of assembly could be selected depending on cell type, interactions with the environment, or state of growth.

Further experiments have shown that the precise shape of the containers is often unimportant for the choice of final assembled structures. For instance, we have regularly observed formation of circular vortices in square chambers (data not shown). However, in a chamber of the same torus-like shape but of a larger lateral size, a different final structure is observed (Fig. 2d).

A surprising variety of larger-scale patterns can be formed in an unconfined geometry by further self-organization of the previously

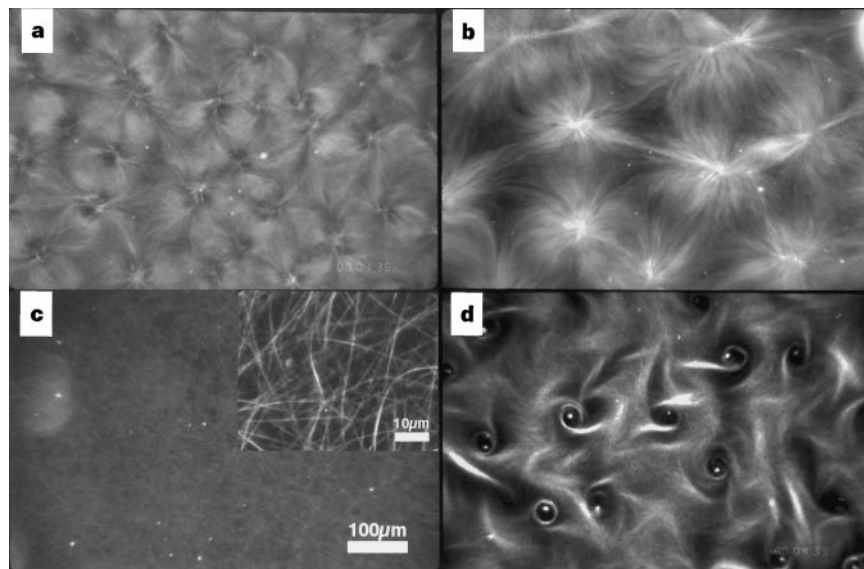


Figure 3 Different large-scale patterns formed through self-organization of microtubules and motors. Initially uniform mixtures of proteins were heated to 37 °C, and patterns resulting after 7 min are shown at equal magnification. The samples differ in kinesin concentration. **a**, A lattice of asters and vortices obtained at $\sim 25 \mu\text{g ml}^{-1}$ kinesin. **b**, An irregular lattice of asters obtained at $\sim 37.5 \mu\text{g ml}^{-1}$ kinesin. **c**, Microtubules form bundles at $\sim 50 \mu\text{g ml}^{-1}$ kinesin (scale bar, 100 μm). Insert, at higher magnification (scale bar, 10 μm). **d**, A lattice of vortices obtained at $< 15 \mu\text{g ml}^{-1}$ kinesin.

described structural elements, asters and vortices. The final patterns depend on the initial concentrations of molecular components: a twofold variation in the concentration of kinesin, included in the same solution of assembling microtubules, results in quite different patterns (Fig. 3). At low motor concentrations a lattice of vortices forms, but at slightly higher kinesin concentrations lattices of asters can be observed. Although the asters obtained in our simplified systems are similar to those observed in cells or cellular extracts, the precise dynamical parameters characterizing both microtubules and motors are very different. When two neighbouring asters overlap sufficiently, they can merge, and this process may determine the final distance between the asters. Finally, at even higher motor concentrations, a distinct state is achieved in which the microtubules form bundles. This process is sensitive to the microtubule nucleation rate, and depends on the ionic conditions.

We have shown that the dynamics of microtubules and motors has to be controlled for a desired structure, such as an aster or a vortex, to be obtained. In our simplified systems this is achieved by controlling the motor and tubulin concentrations. However, in the cellular context the choice of different structures does not necessarily involve the precise variation of relative protein concentrations, but rather the control of their activities. The cell can control the action of motors and the dynamics of microtubule assembly through various regulatory processes, such as covalent modifications and binding of associated proteins^{21,22}.

The self-organization we have observed can be compared to the pattern formation encountered in simple physical systems²³, in which a variety of structures has been formed far from thermal equilibrium from a homogeneous solution of proteins and from a gel of fibres, connected through non-covalent dynamic crosslinks. The energy dissipated in these structures is injected into the system through ATP and GTP hydrolysis²⁴. From this point of view, the situation is reminiscent of patterns formed in chemical reactions²³. The use of protein components, however, means that there is the possibility of additional control at the molecular level. A complete characterization of the phenomena described here should really include an evaluation of all the relevant time and length scales, and the determination of the out-of-equilibrium phase diagrams.

These simplified experiments show that the basic structural 'vocabulary' used by the cell is extremely rich. By using just two basic components and simple local rules of interaction, we obtained a large variety of assembled structures. By extending our system as to include other interacting components, such as nucleation centres for microtubules or different motors, it will be possible to explore the conditions and the components needed for the formation other

structures. Another direction of study would be to introduce some elements of regulation, making it possible to search for rules underlying the choice of different 'words' from this large 'vocabulary' of self-organized structures. □

Methods

Proteins. Recombinant kinesin K401-bio, consisting of 401 amino acids of the N-terminal motor domain of the heavy chain of *D. melanogaster* kinesin linked to the BCCP sequence of *Escherichia coli*, was expressed in *E. coli* and purified²⁵. These recombinant kinesins form biotinylated active dimers²⁵, which can then be joined by adding streptavidin²⁶ (Molecular Probes). A solution of motor constructs was obtained by mixing streptavidin and purified K401-bio, giving final concentrations of 30 $\mu\text{g ml}^{-1}$ and $\sim 150 \mu\text{g ml}^{-1}$, respectively. Complex formation was verified by gel filtration and light scattering. Tubulin was purified from bovine brain⁶; the stock solution was at $\sim 15.3 \text{ mg ml}^{-1}$ in storage buffer (80 mM PIPES and KOH to pH 6.8, 2 mM MgCl_2 , 1 mM EGTA).

Assay with stabilized microtubules. To polymerize microtubules, a solution containing 1.25 mg ml^{-1} tubulin, 1 mM GTP, 1% DMSP in storage buffer was kept at 37 °C for 15 min, and taxol was then added to a final concentration of 20 μM . Polymerized microtubules were kept at room temperature. To start self-assembly, polymerized microtubules, motor constructs and ATP solution were mixed to give final concentrations of 0.27 mg ml^{-1} tubulin, 50 $\mu\text{g ml}^{-1}$ K401-bio, 10 $\mu\text{g ml}^{-1}$ Streptavidin, 1.4 mM ATP, 0.6 mM GTP, 7.7 mM MgCl_2 , 4.3 mM KCl, 1 mM EGTA, 27 $\mu\text{g ml}^{-1}$ α -casein (Sigma), 0.3% DMSO, 20 μM taxol, 80 mM PIPES/KOH to pH 6.8. This mixture was immediately examined by microscopy. Kinesin K401-bio moves microtubules with typical speeds up to $\sim 0.8 \mu\text{m s}^{-1}$ in motility assays (data not shown). We have varied the relative amount of streptavidin to kinesin and found that the maximum activity in motility assays is observed at a stoichiometric ratio of approximately four K401-bio dimers per streptavidin tetramer; this ratio was used for all experiments. The precise ratio of active kinesins per streptavidin might vary to some extent in our experiments owing to some variability of the kinesin activity of different purifications. All experiments were performed several times and the resulting structures were found to be reproducible. No asters are formed without streptavidin or with too much streptavidin. In the absence of ATP, a crosslinked gel of microtubules forms not showing any further temporal evolution.

Assay with non-stabilized, dynamic microtubules. A mixture of purified (unpolymerized) tubulin, motor constructs and nucleotides was now used. Apart from an increased tubulin concentration of 5.1 mg ml^{-1} and absence of taxol, all other concentrations were kept as in the assay with stabilized microtubules, unless specified otherwise. Polymerization of microtubules was started by heating the sample to 37 °C on the microscope.

Glass cleaning and coating. VWR-brand 24 $\times$ 60 mm no. 1 coverslips were cleaned by three rounds of sonication in a hot solution ($\sim 80^\circ\text{C}$) of 10% detergent VWR-Extran 1000 in de-ionized water, followed by 5 rounds of

sonication in pure, hot, de-ionized water. Slides were then immersed in pure ethanol and air-dried. Glass cleaned this way was hydrophilic. To prevent kinesin and microtubules from sticking to the surface, slides were coated by dipping them into a solution of 0.1% agarose in de-ionized water and were allowed to dry. This layer was further coated by dipping the glass first into a solution of 0.2% bovine serum albumin in storage buffer filtered at 0.2 μm , and then four times into de-ionized water. Coated slides were kept humid and were used within a few hours.

Fabrication of microchambers. This was performed as described previously⁶. Glass with microfabricated chambers was cleaned and coated as usual. Sealing of the microchambers was achieved by applying a steady $\sim 20 \text{ kg cm}^{-2}$ pressure for 3 min on the pit-containing coverslip and the slide.

Microscopy and imaging. A Zeiss Axiovert 135 TV with an Olympus 100 $\times$ iris objective, an Olympus 20 $\times$ objective and a Zeiss dark-field ultra condenser was used. Both condenser and objective were heated to warm the immersion oil and the sample to 37 $^{\circ}\text{C}$ by circulating warm water through copper tubing wrapped around them. A CCD camera (Paultek) was used to record on S-VHS video tape, and no video processing was necessary to observe microtubules. A Nikon camera was also mounted on the microscope to take slides with Kodak Ektachrome P1600 film.

Simulations. Simulations will be described in detail elsewhere (F.N. *et al.*, manuscript in preparation). Briefly, microtubules were represented by short rigid rods of 6 μm connected with flexible links: they could grow and shrink according to a simple model of dynamic instability²⁷. The simulations were two-dimensional; this corresponds well to experiments performed in thin chambers or between closely spaced glass surfaces, in which microtubules are nearly parallel to the plane of the sample. Motor complexes could bind up to two microtubules. Unbound complexes could diffuse freely, whereas complexes bound to two microtubules exerted a spring-like force between them. The movement of the microtubules was then calculated at each time step by assuming a completely damped viscous regime. At each time step, motors moved along the microtubules with a speed that is a function of the force they exert. Relevant parameters are known from previous experiments, such as a polymerization speed of 2 $\mu\text{m min}^{-1}$ (ref. 28), a microtubule persistence length of 5 mm (ref. 29), a linear force–speed curve with maximum speed 0.8 $\mu\text{m s}^{-1}$, and a maximum force of 5 pN (refs 10, 11). A typical simulation covered 10 min of real time.

Received 19 March; accepted 13 June 1997.

- Kirschner, M. & Mitchison, T. Beyond self-assembly: From microtubules to morphogenesis. *Cell* **45**, 329–342 (1986).
- Howard, J. The movement of kinesin along microtubules. *Annu. Rev. Physiol.* **58**, 703–729 (1996).
- Hyman, A. & Karsenti, E. Morphogenetic properties of microtubules and mitotic spindle assembly. *Cell* **84**, 401–410 (1996).
- Barton, N. R. & Goldstein, L. S. B. Going mobile: Microtubule motors and chromosome segregation. *Proc. Natl Acad. Sci. USA* **93**, 1735–1742 (1996).
- Urrutia, R., McNiven, M., Albanesi, J., Murphy, D. & Kachar, B. Purified kinesin promotes vesicle motility and induces active sliding between microtubules in vitro. *Proc. Natl Acad. Sci. USA* **88**, 6701–6705 (1991).
- Berliner, E. *et al.* Microtubule movement by a biotinylated kinesin bound to a streptavidin-coated surface. *J. Biol. Chem.* **269**, 8610–8615 (1994).
- Holy, T., Dogterom, M., Yurke, B. & Leibler, S. Assembly and positioning of microtubule asters in microfabricated chambers. *Proc. Natl Acad. Sci. USA* **94**, 6228–6233 (1997).
- Nicklas, R. B. & Ward, S. C. Elements of error correction in mitosis: microtubule capture, release, and tension. *J. Cell Biol.* **126**, 1241–1253 (1994).
- Heald, R. *et al.* Self-organization of microtubules into bipolar spindles around artificial chromosomes in *Xenopus* egg extracts. *Nature* **382**, 420–425 (1996).
- Savin, K. E. & Mitchison, T. Mitotic spindle assembly by two different pathways in vitro. *J. Cell Biol.* **112**, 925–940 (1991).
- Hunt, A. J., Gittes, F. & Howard, J. The force exerted by a single molecule against a viscous load. *Biophys. J.* **67**, 766–781 (1994).
- Svoboda, K. & Block, S. M. Force and velocity measured for single kinesin molecules. *Cell* **77**, 773–784 (1994).
- Mayhofer, E. & Howard, J. The force generated by a single kinesin molecule against an elastic load. *Proc. Natl Acad. Sci. USA* **92**, 574–578 (1995).
- Kashina, A. S. *et al.* A bipolar kinesin. *Nature* **379**, 270–272 (1996).
- Arnal, I. & Wade, R. How does taxol stabilize microtubules? *Curr. Biol.* **5**, 900–908 (1995).
- Verde, F., Berrez, J.-M., Antony, C. & Karsenti, E. Taxol-induced microtubule asters in mitotic extracts of *Xenopus* eggs: Requirement for phosphorylated factors and cytoplasmic dynein. *J. Cell Biol.* **112**, 1177–1187 (1991).
- Stearns, T. & Kirschner, M. In vitro reconstitution of centrosome assembly and function: The central role of gamma-tubulin. *Cell* **76**, 623–637 (1994).
- Mitchison, T. & Kirschner, M. Microtubule assembly nucleated by isolated centrosomes and dynamic instability of microtubule growth. *Nature* **312**, 232–242 (1984).
- Theurkauf, W. E. Premature microtubule-dependent cytoplasmic streaming in cappuccino and spire mutant oocytes. *Science* **265**, 2093–2096 (1994).
- Rodionov, V. & Borisy, G. G. Self-centring activity of cytoplasm. *Nature* **386**, 170–173 (1997).
- Vernos, I. & Karsenti, E. Motors involved in spindle assembly and chromosome segregation. *Curr. Opin. Cell Biol.* **8**, 4–9 (1996).

- McNally, F. J. Modulation of microtubule dynamics during the cell cycle. *Curr. Opin. Cell Biol.* **8**, 23–29 (1996).
- Cross, M. C. & Hohenberg, P. C. Pattern formation outside of equilibrium. *Rev. Mod. Phys.* **65**(2) (1993).
- Tabony, J. Morphological bifurcations involving reaction–diffusion processes during microtubule formation. *Science* **264**, 245–248 (1994).
- Young, E. C., Berliner, E., Mahtani, H., Perez-Ramirez, B. & Gelles, J. Subunit interactions in dimeric kinesin heavy chain derivatives that lack the kinesin rod. *J. Biol. Chem.* **270**, 3926–3931 (1995).
- Bayer, E. & Wilchek, M. Avidin and streptavidin. *Methods Enzymol.* **184**, 51–67 (1990).
- Dogterom, M., Leibler, S. Physical aspects of the growth and regulation of microtubule structures. *Phys. Rev. Lett.* **70**, 1347–1350 (1993).
- Erickson, H. & O'Brien, E. Microtubule dynamic instability and GTP hydrolysis. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 145–166 (1992).
- Gittes, F., Mickey, B., Nettleton, J. & Howard, J. Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *J. Cell Biol.* **120**, 923–934 (1993).

Acknowledgements. We thank E. Young and J. Gelles for kinesin plasmids; J. Johnson for taxol; T. Holy for help with the preparation of glass chambers; and T. Holy, M. Elowitz, E. Wolf, E. Karsenti, T. Mitchison, J. Howard and S. Block for discussions. This work was supported by grants from the NIH, the NSF and the HFSP, a fellowship from the Deutsche Forschungsgemeinschaft (T.S.) and the French Government (F.J.N.).

Correspondence and requests for materials should be addressed to S.L. (e-mail: leibler@princeton.edu).

Structure of a viral procapsid with molecular scaffolding

Terje Dokland*, Robert McKenna*, Leodevico L. Ilag*, Brian R. Bowman*, Nino L. Incardona†, Bentley A. Fane§ & Michael G. Rossmann*

* Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392, USA

† Department of Microbiology and Immunology, University of Tennessee, Memphis, Tennessee 38163, USA

§ Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas 72701, USA

The assembly of a macromolecular structure proceeds along an ordered morphogenetic pathway, and is accomplished by the switching of proteins between discrete conformations as they are added to the nascent assembly^{1–3}. Scaffolding proteins often play a catalytic role in the assembly process^{1,2,4}, rather like molecular chaperones⁵. Although macromolecular assembly processes are fundamental to all biological systems, they have been characterized most thoroughly in viral systems, such as the icosahedral *Escherichia coli* bacteriophage Φ X174 (refs 6, 7). The Φ X174 virion contains the proteins F, G, H and J^{7,8}. During assembly, two scaffolding proteins B and D are required for the formation of a 108S, 360-Å-diameter procapsid from pentameric precursors containing the F, G and H proteins^{6,9}. The procapsid contains 240 copies of protein D, forming an external scaffold, and 60 copies each of the internal scaffolding protein B, the capsid protein F, and the spike protein G^{9,10}. Maturation involves packaging of DNA and J proteins and loss of protein B, producing a 132S intermediate^{6,7}. Subsequent removal of the external scaffold yields the mature virion. Both the F and G proteins have the eight-stranded antiparallel β -sandwich motif^{8,11} common to many plant and animal viruses^{12,13}. Here we describe the structure of a procapsid-like particle at 3.5-Å resolution, showing how the scaffolding proteins coordinate assembly of the virus by interactions with the F and G proteins, and showing that the F protein undergoes conformational changes during capsid maturation.

For the structure determination we obtained a low-resolution, 6.5-Å data set collected from a single, frozen crystal, and a higher-resolution, 3.5-Å data set collected from a number of crystals at 4 $^{\circ}\text{C}$ (Table 1). The structure was solved to nearly 3.5-Å resolution by molecular-replacement real-space averaging¹⁴, using a starting model based on the previously determined electron-microscopic

† Present addresses: Department of Biological Sciences, Warwick University, Coventry CV4 7AL, UK (R.M.); Morphosys GmbH, Frankfurter Ring 193a, 80807 München, Germany (L.L.); Department of Biochemistry, Baylor College of Medicine, Houston, Texas 77030, USA (B.R.B.).

reconstruction¹⁰ (Table 2). The procapsid has a spherical appearance, the most prominent features being 12 pentameric rings of density on the 5-fold axes and 30 rhomboidal densities on the 2-fold axes (Fig. 1a). We refer to this particle as the 'closed procapsid' to contrast it with the more open appearance of the previously described electron microscopic structure¹⁰ (the 'open procapsid'), which has 30-Å-wide holes at the 3-fold axes. The 2-fold rhomboidal densities in the closed procapsid are due to the external scaffolding protein D, and are absent from the virion^{8,11} (Fig. 1b). The inner part of the shell and the protruding 5-fold spikes are similar in the virion and closed procapsid structures. The G-protein pentameric spike is translated radially outwards by 5.5 Å along the 5-fold axis in the closed procapsid relative to the virion. Similarly, the F-protein β -sandwich domain (residues 10–175, 233–304 and 320–413) of the closed procapsid is translated outwards by 3.3 Å, leaving the 5-fold interactions unchanged (Fig. 2). An α -helical domain, consisting of residues 175–207 and 305–319, is in almost the same position as in the virion, thus maintaining the 3-fold subunit contacts. The result is a hinge movement of 8.3° between the β -sandwich domain and the α -helical domain. The largest con-

formational differences occur between these two domains, from residues 208 to 232 (Fig. 2).

Comparison of the X-ray structure of the mature virion with the electron-microscopic reconstruction of the open procapsid suggested that the F capsid protein undergoes major conformational changes during maturation¹⁰, as seen in many double-stranded DNA viruses^{15,16}. DNA packaging would be expected to take place before the conformational change of the F protein, while the shell still has openings around the 3-fold axes for entry of the genome¹⁰. The relatively small differences between the virion and the closed procapsid (Fig. 2) may therefore be a relic of much greater differences occurring at an earlier stage of assembly. The conformational changes that produced the closed procapsid found in the crystal structure could have been triggered by the conditions used for the crystallization of the particles. Viral procapsids are metastable particles in which spontaneous conformational changes are often observed. The same situation occurred in the case of the 'degraded procapsid' of G4, a phage related to Φ X174 (ref. 17). Although the closed procapsid might be an intermediate (perhaps in dynamic equilibrium with the open procapsid) that appears before DNA packaging, it is more likely to be an abortive particle, as it contains no DNA and has no holes large enough for the genome to enter.

There are four copies of the external scaffolding protein D per icosahedral asymmetric unit (D1–D4; Fig. 3a) in the closed procapsid structure, giving a total of 240 copies in the entire shell. Each copy of D has a predominantly α -helical fold. There are seven principal helical stretches (α 1– α 7), with mostly short loop regions in between (loops 1–6) (Fig. 3a). Loop 5 also includes some β -structure and has different conformations in each of the four subunits. Subunits D2, D3 and D4 are lined up along the base of the icosahedral asymmetric unit triangle, between the two 3-fold axes (Figs 1a, 3a), and make contacts with their three 2-fold symmetry-related neighbours (Fig. 1a, Table 3). The D1 subunit is nearer the 5-fold axis at the top of the triangle, where it interacts with the G proteins in the pentameric spike. The D proteins form a fenestrated scaffolding shell that makes few contacts with the underlying capsid (Fig. 3b, Table 3). The four copies of protein D are not related by any closed point group symmetry; rather, each subunit has a radically different environment, so that very different parts of the molecule make contact with neighbouring proteins. This organization bears no relation to a classical $T = 4$ arrangement, in which differences in contacts would be minimized by placing the subunits in quasi-equivalent positions¹⁸. One explanation

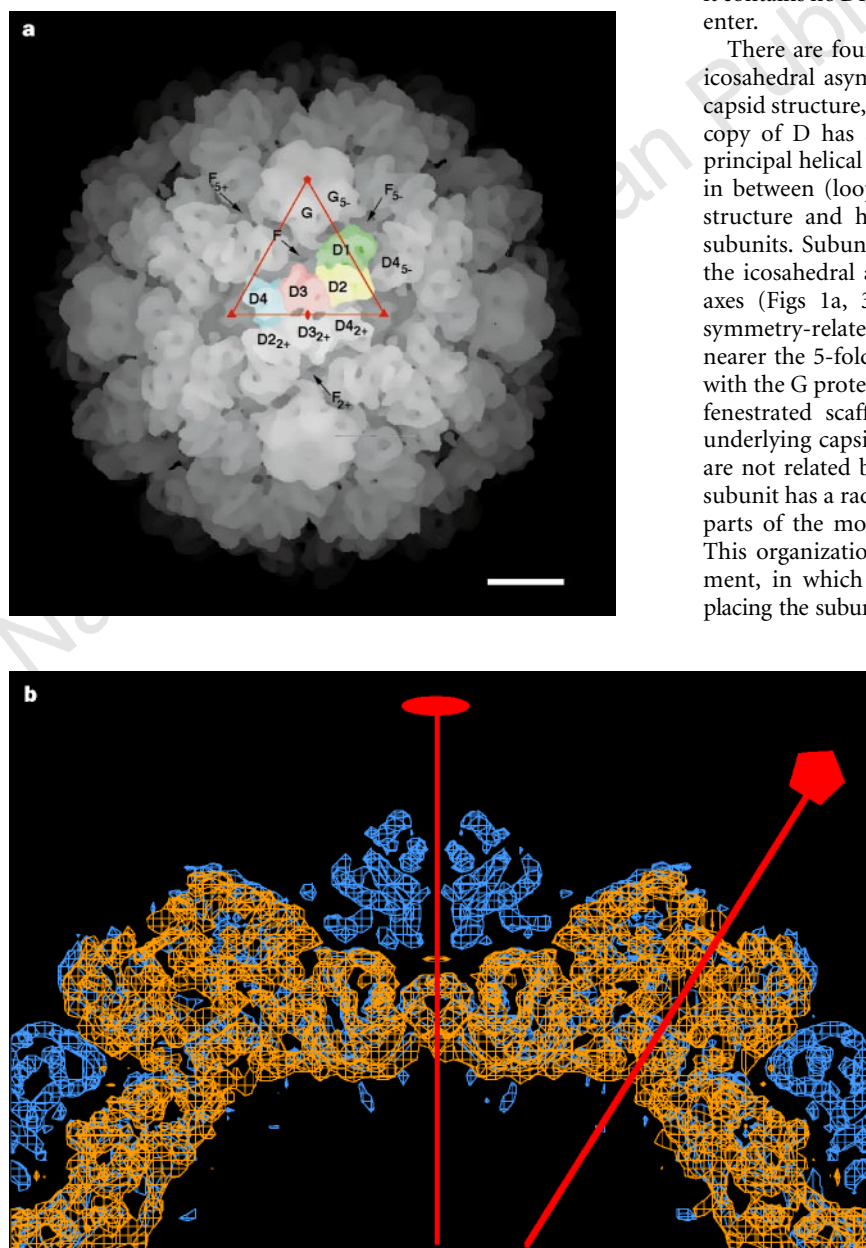


Figure 1 **a**, Isosurface representation of a 10-Å resolution electron-density map of the whole closed procapsid particle, viewed down a 2-fold axis. One asymmetric unit triangle, delimited by a 5-fold, a 2-fold and two 3-fold axes, is shown in red. The four D proteins, D1 (green), D2 (yellow), D3 (red) and D4 (blue), in the asymmetric unit are identified, as are the G proteins in the 5-fold spike and the underlying F proteins. Symmetry-related neighbouring subunits are indicated by the subscripts 5- and 5+ for 5-fold related subunits and 2+ for 2-fold related subunits. Scale bar, 50 Å. **b**, A 5-Å-thick slab through the centre of the virion electron-density map⁸ (orange), superimposed on the corresponding closed procapsid map (blue), at 6.5-Å resolution. The positions of one 2-fold and one 5-fold axis are indicated. The additional external density around the 2-fold axis in the closed procapsid map is due to the D scaffolding protein.

for this lack of local symmetry is that the scaffold is coordinated by the underlying $T = 1$ shell through interactions with the F and G proteins. The D protein has a structurally conserved core, amounting to 75 residues or about 50% of the sequence, for which the C α atoms can be superimposed with a root-mean-squared (r.m.s.) deviation of less than 2.5 Å between any pair of subunits (Fig. 3a).

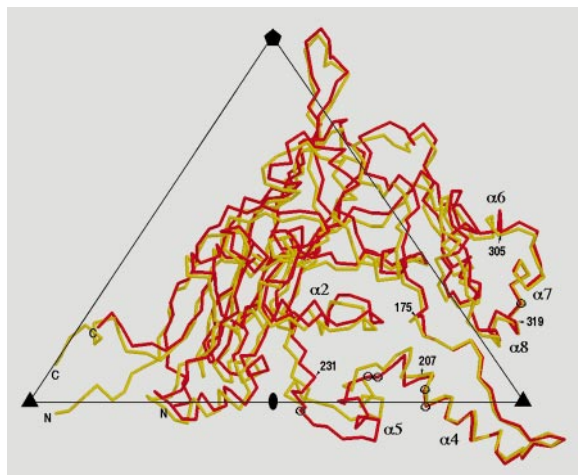


Figure 2 C α backbone plot of the F protein in the closed procapsid (red). The asymmetric unit triangle is shown, with the symmetry axes marked. The F-protein structure from the virion (yellow) was superimposed on the closed procapsid F protein by minimizing the r.m.s. distance between corresponding C α atoms in the α -helical domain (residues 175–207, 305–319). The main difference in the F protein between the virion and the closed procapsid is a hinge movement between the α -helical domain and the β -sandwich domain. Residues involved in second-site suppression of cold-sensitive D mutants²¹ are circled.

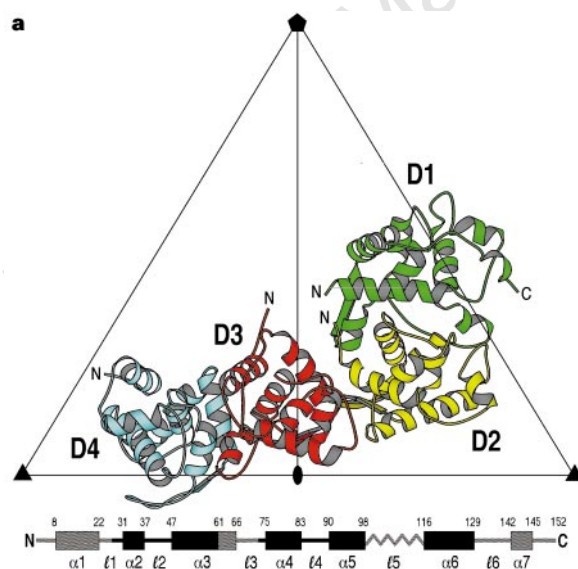
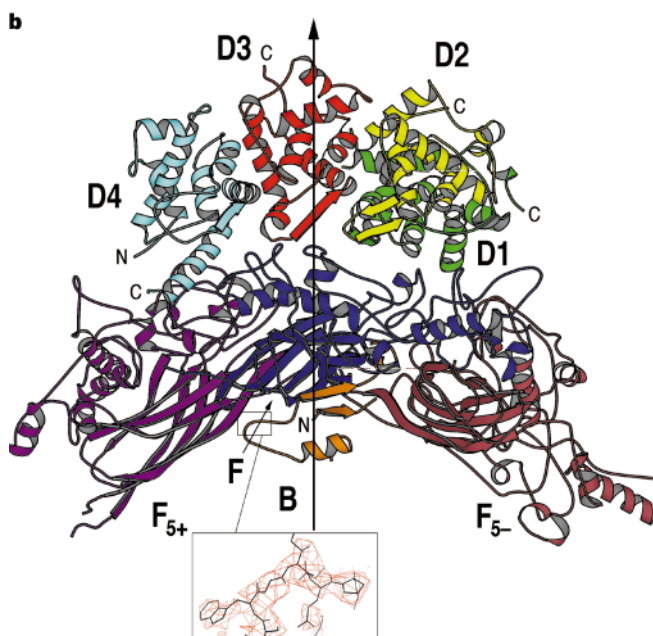


Figure 3 a, Ribbon diagram²⁹ showing the arrangement of the four D scaffolding proteins in one icosahedral asymmetric unit triangle, coloured as in Fig. 1a, viewed down a 2-fold axis. The symmetry axes and the visible N and C termini are indicated. The secondary structure assignment is indicated on the line below. Boxes represent the α -helical regions (α 1– α 7); connecting loops are labelled l1 to l6. The hatched part of the sequence represents structurally variable regions, whereas the solid part represents the structurally conserved core. The zigzag line

The main regions of conformational variability among the four D subunits include the amino and carboxy termini (residues 1–26 and 130–152), as well as loop 5 (residues 100–115) and residues 61–72, which include part of α 3 and loop 3 (Figs 3a, 4). Subunits D1 and D3 are in very similar conformations, except for a few residues at the termini. These two subunits have a $\sim 30^\circ$ kink in α 3 at Gly 61, and loop 5 has a β -strand with a small α -helix embedded in its tip. Similarly, subunits D2 and D4 are similar; in these two subunits α 3 lacks the kink and loop 3 is narrower than in D1 and D3. There are, however, some important differences between D2 and D4. Loop 5 in D2 has an antiparallel β -hairpin conformation, whereas in D4 this loop is much wider and stands out further from the body of the protein. The terminal helices in D4 (α 1 and α 7) have undergone dramatic changes in their direction relative to the other subunits (Figs 3b, 4). Helix α 7 forms the strongest interactions with the F capsid protein. The partial or complete disorder of this helix in the other subunits suggests that the orientation of α 7 is fixed only when it can find a suitable anchor.

Although there is no closed point group symmetry relating the four independent D subunits, they are organized into two dimers (D1 : D2 and D3 : D4), within which the monomers are related by a 56° rotation. The contacts between the monomers in each dimer are similar. Of the 304 C α atoms in each dimer, 67% superimpose with an r.m.s. deviation of less than 2.5 Å (Figs 3a, 4). Loop 3 of D1 and D3 makes mainly hydrophobic contacts with α 5 and α 6 in D2 and D4, respectively. In D2 and D4, loop 3 is in a different environment, which explains its different conformation. Loop 3 is rich in conserved proline and glycine residues, which are typically involved in structural flexibility and hinge motion. The interactions between the D1 : D2 and D3 : D4 dimers, however, involve mainly charged residues in α 1 and α 4. The relatively few contacts across the icosahedral 2-fold interface are made predominantly between loop 5 of D4 and α 3 of D2, and between the α 3/loop 3 regions of



represents the variable loop 5, which contains some β structure. **b**, A side view, perpendicular to that in **a**, showing the four copies of the D protein, the underlying F protein (purple), and its 5-fold related neighbours F_{5+} (fuchsia) and F_{5-} (mauve), and the internal B scaffolding protein (orange). The predominant interaction between the D and F proteins is made between the C-terminal helix α 7 of D4 and the F_{5+} subunit (Table 3). A 2-fold axis is pointing straight up. Inset, part of the B protein and corresponding density.

2-fold related D3 subunits. There are few contacts between D4 and the adjacent D1 and D2 subunits in the 5-fold related asymmetric unit (Table 3).

In the closed procapsid there is additional density caused by protein B on the inner surface of the F capsid shell. Although this density is poorly defined, the C-terminal 39 residues of the B protein could be built with reasonable accuracy. Residues 84–91 form an α -helix, whereas residues 105–109 have a β -strand conformation. Density corresponding to an additional strand of about 10 residues ending near the 2-fold axis may represent the N-terminal part of the protein (Fig. 3b). The remaining half of the protein cannot be identified in the present map. Previous work based on cryo-electron-microscopic reconstructions had suggested that the B protein was situated on the inside of the capsid shell at the 2-fold axes¹⁰, which corresponds well with the location occupied by the B-protein density in the crystal structure. Presumably, the B proteins make dimeric contacts across the 2-fold axes. The C-terminal part of protein B fits into a pocket made between $\alpha 2$ and the β -sandwich of the F protein, including residues Lys 166, Arg 239 and Arg 290, of which Arg 290 interacts with the C terminus of protein B. The binding site on the F protein for the C-terminal part

of protein B overlaps with that of the C-terminal part of the J protein in the virion¹¹. This explains the unexpected finding of electron density representing protein J in a previous study on the 'degraded procapsid' of the related phage G4 (ref. 17). Most probably, that density should have been interpreted as the C terminus of protein B, which has similar aromatic characteristics to protein J in that region. For proteins B and J to share an interaction site on protein F may be functionally important, because the J protein enters the capsid as the B protein leaves during DNA packaging^{6,7}, suggesting that there may be competition between the B and J proteins for the same binding site.

The dimeric subassembly units (D1:D2 and D3:D4) seen in the structure of the closed procapsid probably represent a basic assembly module of the scaffold. The existence of the same dimer in different environments suggests that this is a stable subassembly and forms the initial morphogenetically relevant scaffold precursor (Fig. 5a). This dimer may be a component of the tetramer found in solution¹⁹, although, as the dimer does not form a closed point group, what prevents the assembly from using the same interactions to grow indefinitely into a helix¹⁸? Presumably, the second monomer is switched into a non-sticky conformation after binding the first

Table 1 Data processing statistics

Data set	No. of crystals	Resolution (Å)	R_{merge}^* ($R_{\text{merge},2\sigma}$)	Completeness† (multiplicity)	No. of observed (unique) reflections	No. of measured reflections
Frozen	1	6.5	15.0 (6.7)	97.0 (4.6)	646,058 (141,534)	1,906,599
Warm‡	22	3.5	24.7 (12.9)	55.3 (2.0)	1,072,459 (527,445)	1,646,040

* $R_{\text{merge}} = 100 \times \sum_i \sum_j |I_{ij} - \langle I_{ij} \rangle| / \sum_i \sum_j I_{ij}$; $R_{\text{merge},2\sigma}$ is calculated using only data for which $F^2 > 2 \times \sigma(F^2)$.

† Completeness = $100 \times (\text{no. of observed unique reflections}) / (\text{no. of possible unique reflections})$; multiplicity = $(\text{no. of observed reflections}) / (\text{no. of observed unique reflections})$.

‡ Includes scaled partial reflections with partiality >0.5.

Table 2 Real-space averaging statistics

Data set	Averaging cycle no	Particle positions*					Resolution (Å)	CC†	R‡
		1		2					
		κ	χ	κ	χ				
Frozen, one particles	1	0.00	0.2450				13.0	0.3919	53.1
	31	0.00	0.2488				13.0	0.5970	42.6
Frozen, two particles	1	0.00	0.2448	−43.40	−0.0050		13.0	0.5823	43.8
	487	−0.63	0.2446	−42.68	−0.0046		6.5	0.9560	13.2
Warm, two particles	1	−0.63	0.2446	−42.68	−0.0046		9.0	0.6770	35.1
	305	−0.04	0.2432	−41.31	−0.0052		4.0	0.9274	16.4

* χ is the position of the particle in fractional unit cell coordinates ($\chi = y = z$, because the particle sits on a 3-fold axis along the body diagonal of the unit cell); κ is the rotation of the particle around this 3-fold axis starting from the standard icosahedral orientation for which the icosahedral 2-fold axes are parallel to the crystallographic axes. For definition of the polar coordinates, see ref. 28. (A rotation of $\kappa = 44.48^\circ$ around the 3-fold axis corresponds to a 90° rotation around the 2-fold axis.)

† $\text{CC} = [\sum_i (F_{\text{obs},i} - F_{\text{calc},i})^2 / \sum_i (F_{\text{obs},i} - F_{\text{calc},i})^2 + \sum_i (F_{\text{calc},i} - F_{\text{obs},i})^2 / \sum_i (F_{\text{calc},i} - F_{\text{obs},i})^2]^{1/2}$.

‡ $R = 100 \times \sum_i |F_{\text{calc},i} - k F_{\text{obs},i}| / \sum_i |F_{\text{calc},i}|$, where $k = (F_{\text{calc},i} / F_{\text{obs},i})$ for all reflections within each resolution bin.

§ Initially, only the particle at (1/4,1/4,1/4) was used for phasing. The particle at (0,0,0) was introduced after cycle 31.

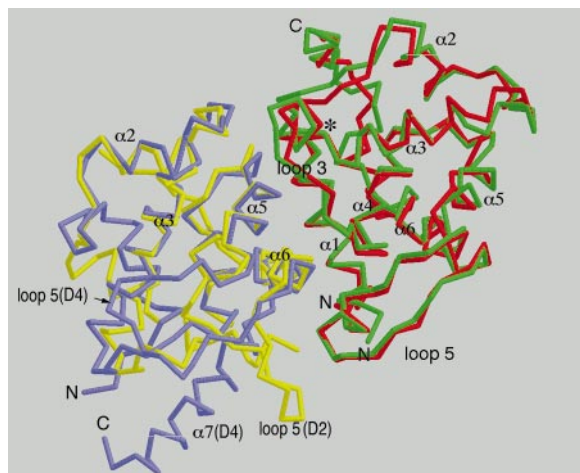


Figure 4 Structural comparison of the D subunits. Only the C α backbones are shown. Superposition of the D1:D2 (green:yellow) dimer onto the D3:D4 (red:blue) dimer. (The colour scheme is as in Fig. 3.) The superposition was achieved by minimizing the r.m.s. distance between corresponding C α atoms within the structurally conserved core region. Secondary structure elements are indicated. The asterisk represents the kink in $\alpha 3$ of subunits D1 and D3.

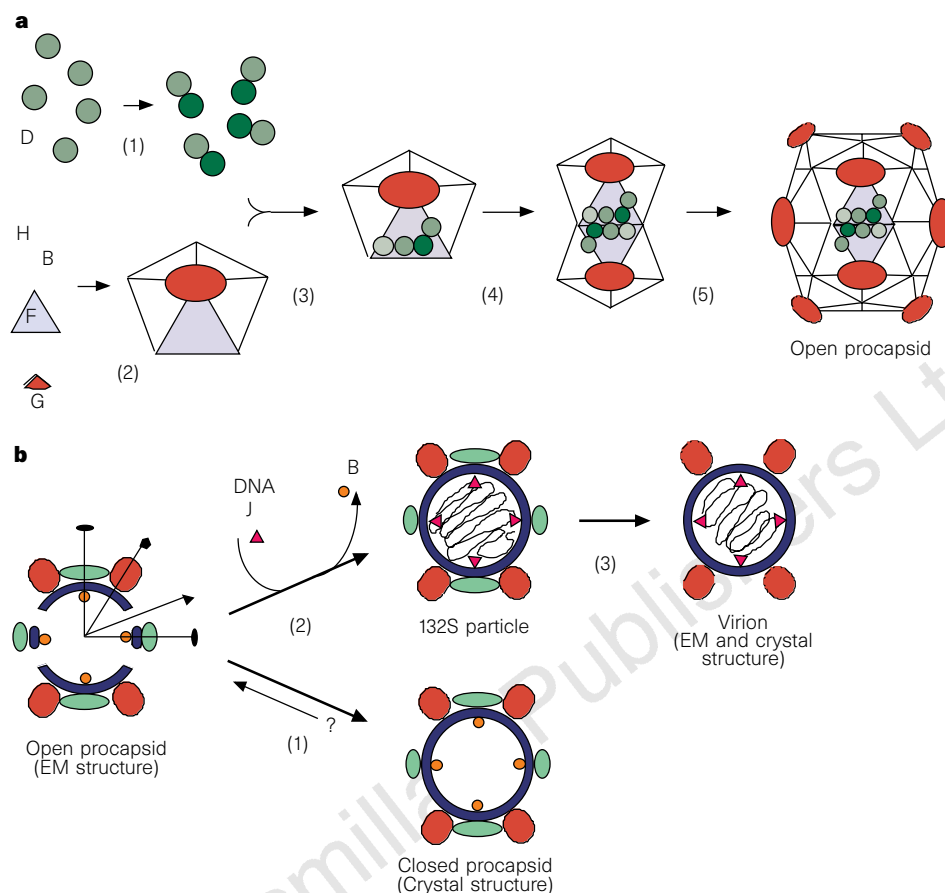


Figure 5 a, Early steps in the assembly pathway of ϕ X174. External view, depicting the F capsid protein as triangles, pentamers of the G spike protein as red ellipsoids, and the D scaffolding protein as green circles. The different shading patterns in the D proteins represent the different conformational states. (1) The D scaffolding protein forms asymmetric dimers. (2) The main capsid protein F forms pentameric precursors, probably complexed with B, G and H proteins. One of the five asymmetric units is shaded blue. (3) The D dimers bind to the pentameric precursors in at least two separate steps, conferring a unique conformation on one of the D subunits (very light green). (4) Binding to F enables the D dimer to bind another dimer across the 2-fold interface. The B protein (not shown, on the inside) is also required for this reaction. (5) Generation of these 2-fold contacts leads to the formation of a complete icosahedral procapsid shell. **b**, Maturation of

the ϕ X174 open procapsid. The particles are shown in cross-sections, but the colour scheme is as in **a**. Protein B is depicted in orange. One 5-fold and two 2-fold axes are indicated on the open procapsid. The open procapsid, for which the structure was determined by electron microscopy (EM), has holes on the 3-fold axes. (1) The closed procapsid seen in the crystal structure may be an abortive by-product, or be in equilibrium with the open procapsid. It has undergone conformational changes relative to the open procapsid, and its F and G proteins are similar to those of the virion. (2) Unlike the closed procapsid, the 132S particle has undergone packaging of DNA and J protein, but has lost the B protein. It is probably structurally similar to the closed procapsid. (3) Removal of the external scaffolding protein D from the 132S particle generates the mature, infectious virion.

monomer. The straightening of $\alpha 3$ and corresponding conformational change of loop 3 between the two subunits of the dimer are probably involved in this switch. The first interactions between the D and F proteins are probably formed by assembly-naïve D dimers binding to pentameric precursors made of F and G proteins (Fig. 5a). There would be no need for strong 5-fold contacts within the D scaffold, as this interaction is already organized in the preformed underlying F pentamers. Nevertheless, there is very little contact between the D and F proteins in the closed procapsid. These contacts are made almost entirely through the conformationally unique $\alpha 7$ of D4 (Fig. 3b). Genetic analysis, however, showed that cold-sensitive mutations (Q12W and Q22W) within $\alpha 1$ of the D protein confer a 'fragile procapsid' phenotype, characterized by dissociation of the shell during DNA packaging. This phenotype can be suppressed by various mutations in or near $\alpha 4$ of the F protein^{20,21} (Fig. 2, circled), indicating that there is a more intimate association between $\alpha 4$ of protein F and $\alpha 1$ of protein D at an earlier stage in morphogenesis. This interaction could keep the holes at the 3-fold axes open until DNA packaging is completed.

Both the D and B scaffolding proteins are required for the procapsid to be formed^{6,7}. In their presence, preformed pentamers

containing proteins F, G and H assemble into a shell^{6,7} by means of contacts across the icosahedral 2-fold axes (Fig. 5a). Because the pentamers are incapable of forming these interactions by themselves, it follows that the B and D proteins are involved in making these contacts. The most prominent 2-fold interactions in the external scaffold are made by the conformationally unique loop 5 of D4, which reaches across the 2-fold interface to make contact with D2 in the 2-fold related asymmetric unit. The initial binding of D4 to F could switch the structure of loop 5 into an assembly-competent form, facilitating the binding of D4 to D2. Similarly, 2-fold contacts are made on the inside of the shell through interactions between the B proteins. Once these 2-fold contacts have been made by the scaffolding proteins, the closing of the 3-fold holes, resulting from extensive interactions between the surrounding helical domains of the F proteins, will consolidate the assembly of the virion and allow the removal of the scaffolding proteins (Fig. 5). □

Methods

Purification and crystallization. The ϕ X174 procapsids were produced from lysis-defective mutant infections of wild-type cells^{9,10} and mutant cells that

Table 3 Contacts between proteins

Protein	D1	D2	D3	D4	B
D1		69			
D2	69		123		
D3		123		73	
D4			73		
D4 ₅₊	12	9			
D2 ₂₊				32	
D3 ₂₊			54		
F	5	4	24	9	158
F ₅₊				39	
F ₂₊				5	
G	12				
G ₅₊	24				

The number of total interatomic contacts (interatomic distance <3.6 Å) made between pairs of proteins in the closed procapsid structure are shown. The subscript 5+ indicates a subunit related by a +72° rotation around the 5-fold axis at the vertex of one asymmetric unit, and 5- represents a -72° rotation. Subscript 2+ denotes a subunit related by a 180° rotation around the 2-fold axis. Refer to Fig. 1a to identify the interactions.

arrest the assembly before DNA packaging²². After purifying twice on 10–30% tartrate gradients (originally 5–25% sucrose), the suspension was made to a concentration of 15–20 mg ml⁻¹ in a buffer containing 10 mM Tris, pH 7.5, 0.25 M NaCl, 5 mM EDTA and 5 mM MgCl₂. The particles were crystallized by hanging-drop vapour diffusion against 42–47% saturated ammonium sulphate in 0.1 M MES, pH 6.0. For freezing, the crystals were soaked for 30 min in 20% glycerol in a stabilizing solution (47–50% ammonium sulphate in 0.1 M MES, pH 6.0), and pre-frozen in liquid ethane.

Data processing. The frozen crystals never diffracted to better than 6.5-Å resolution; higher-resolution data (3.5-Å) were collected at 4 °C ('warm' data). Only 2–4 images could be collected per crystal, owing to the radiation sensitivity of the warm crystals, but a complete data set to 6.5-Å resolution was collected from a single, frozen crystal. The data sets were collected on Fuji image plates and on film at the Cornell High Energy Synchrotron Source (CHESS) F-1 beam line and on a MAR 30-cm image-plate scanner at the Brookhaven beam line X12. Oscillation angles were either 0.25° or 0.3°. Images were processed using DENZO software²³, and the frozen data were scaled, postrefined and merged using SCALEPACK²³. The frozen data had a mosaic spread of 0.26°, whereas the mosaic spread for the warm data varied from 0.08° to 0.29°. The warm data contained a high proportion of unpaired partial reflections. SCALA and AGROVATA from the CCP4 suite²⁴ were used to scale and merge this data set, after postrefining in SCALEPACK. Reflections that extended over more than one image were added, and unpaired reflections with a partiality greater than 0.5 were corrected (Table 1).

Structure determination. The space group of the crystals is *I*2₁3, with *a* = 769 Å (frozen data) or 774 Å (warm data). The *V*_M (Matthews coefficient) is 3.1 Å³ Da⁻¹ assuming 16 particles per unit cell. There are two independent particles per asymmetric unit, each sitting on a 3-fold axis. The resulting close-packed arrangement places the particles near (0,0,0) and (1/4,1/4,1/4), with a closest approach of 333 Å. This requires the determination of one positional and one rotational parameter for each particle. The initial particle orientations were determined from a self-rotation function²⁵, which showed that the two independent particles were related by an approximately 90° rotation around the icosahedral 2-fold axes that were approximately parallel to the crystallographic 2-fold axes. This orientation allowed one of the two particle positions to be determined accurately from the Patterson function. The other particle had a larger rotational offset, so its position could not be determined directly from the Patterson, but was inferred from packing considerations.

A model made by fitting the X-ray structures of the virion proteins⁸ into a 25-Å electron-microscopic reconstruction of the open procapsid¹⁰ was used as a phasing start for the molecular-replacement real-space averaging²⁶ using the more complete frozen data. It turned out that this was a poor model, which essentially represented random atoms within an envelope that was itself only partly correct. The starting resolution for the averaging was 20–13 Å, which was extended one grid point at a time to 6.5 Å. Calculated structure factors were used where there were no observed data. Refinement of the initial positions and orientations moved the particles by up to 4 Å and 1.8°, respectively (Table 2). An averaged electron-density map calculated from the frozen data was then used as a starting model for the phasing of the warm data. After further

refinement of the position and orientation of each of the two particles, the resolution was extended two grid points at a time to 3.5 Å (Table 2). During most of the procedure, the two independent particles were averaged separately using 20-fold non-crystallographic averaging, but 40-fold averaging was introduced only in the last few cycles. The final map was calculated using data from 23–3.5 Å resolution. Considering the low quality and completeness (13.5%) of the data between 3.5- and 4-Å resolution, the map was good. It was possible to recognize the side chains with sufficient confidence to build the amino-acid sequences of the D, F, G and part of the B proteins into the map, using the program O²⁷. Because of the high non-crystallographic symmetry, the phase determination is well restrained. Thus crystallographic refinement would result in only a small improvement in the fit of the structure to the available density, but would have little effect on the quality of the electron-density distribution.

Received 24 March; accepted 10 June 1997.

- King, J. in *Biological Regulation and Development* Vol. 2, *Molecular Organization and Cell Function* (ed. Goldberger, R. E.) 101–132 (Plenum, New York, 1980).
- Casjens, S. & Hendrix, R. in *The Bacteriophages* Vol. 1 (ed. Calendar, R.) 15–75 (Plenum, New York, 1988).
- Osawa, F. & Asakura, S. *Thermodynamics of the Polymerization of Proteins* (Academic, London, 1975).
- Preston, V. G., al-Kobaisi, M. F., McDougall, I. M. & Rixon, F. J. The herpes simplex virus gene UL26 proteinase in the presence of the UL26.5 gene product promotes the formation of scaffold-like structures. *J. Gen. Virol.* **75**, 2355–2366 (1994).
- Martin, J. & Hartl, F. U. Chaperone-assisted protein folding. *Curr. Opin. Struct. Biol.* **7**, 41–52 (1997).
- Hayashi, M. in *The Single-Stranded DNA Phages* (eds Denhardt, D. T., Dressler, D. & Ray, D. S.) 531–547 (Cold Spring Harbor Laboratory Press, NY, 1978).
- Hayashi, M., Aoyama, A., Richardson, D. L. Jr & Hayashi, M. N. in *The Bacteriophages* Vol. 2 (ed. Calendar, R.) 1–71 (Plenum, New York, 1988).
- McKenna, R. *et al.* Atomic structure of single-stranded DNA bacteriophage φX174 and its functional implications. *Nature* **355**, 137–143 (1992).
- Mukai, R., Hamatake, R. K. & Hayashi, M. Isolation and identification of bacteriophage φX174 prohead. *Proc. Natl Acad. Sci. USA* **76**, 4877–4881 (1979).
- Ilag, L. L. *et al.* DNA packaging intermediates of bacteriophage φX174. *Structure* **3**, 353–363 (1995).
- McKenna, R., Ilag, L. L. & Rossmann, M. G. Analysis of the single-stranded DNA bacteriophage φX174, refined at a resolution of 3.0 Å. *J. Mol. Biol.* **237**, 517–543 (1994).
- Harrison, S. C., Skehel, J. J. & Wiley, D. C. in *Fields Virology* Vol. 1 (eds Fields, B. N., Knipe, D. M. & Howley, P. M.) 59–99 (Lippincott-Raven, Philadelphia, 1996).
- Rossmann, M. G. & Johnson, J. E. Icosahedral RNA virus structure. *Annu. Rev. Biochem.* **58**, 533–573 (1989).
- Rossmann, M. G. The molecular replacement method. *Acta Crystallogr. A* **46**, 73–82 (1990).
- Dokland, T. & Muriel, H. Structural transitions during maturation of bacteriophage lambda capsids. *J. Mol. Biol.* **233**, 682–694 (1993).
- Prasad, B. V. *et al.* Three-dimensional transformation of capsids associated with genome packaging in a bacterial virus. *J. Mol. Biol.* **231**, 65–74 (1993).
- McKenna, R., Bowman, B. R., Ilag, L. L., Rossmann, M. G. & Fane, B. A. Atomic structure of the degraded procapsid particle of the bacteriophage G4: induced structural changes in the presence of calcium ions and functional implications. *J. Mol. Biol.* **256**, 736–750 (1996).
- Caspar, D. L. D. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24 (1962).
- Farber, M. B. Purification and properties of bacteriophage φX174 gene D products. *J. Virol.* **17**, 1027–1037 (1976).
- Ekechukwu, M. C. & Fane, B. A. Characterization of the morphogenetic defects conferred by cold-sensitive prohead accessory and scaffolding proteins of φX174. *J. Bacteriol.* **177**, 829–830 (1995).
- Fane, B. A., Shien, S. & Hayashi, M. Second-site suppressors of a cold sensitive external scaffolding protein of bacteriophage φX174. *Genetics* **134**, 1003–1011 (1993).
- Ekechukwu, M. C., Oberste, D. J. & Fane, B. A. Host and φX174 mutations affecting the morphogenesis or stabilization of the 50S complex, a single-stranded DNA synthesizing intermediate. *Genetics* **140**, 1167–1174 (1995).
- Otwinski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276A**, 307–326 (1997).
- Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Tong, L. & Rossmann, M. G. The locked rotation function. *Acta Crystallogr. A* **46**, 783–792 (1990).
- Rossmann, M. G. *et al.* Molecular replacement real-space averaging. *J. Appl. Crystallogr.* **25**, 166–180 (1992).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Rossmann, M. G. & Blow, D. M. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr. B* **15**, 24–31 (1962).
- Kraulis, P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).

Acknowledgements. We thank the staff at the Cornell High Energy Synchrotron Source and the National Light Source staff at beam line X12 for providing the data collection facilities; W. F. Bean for help with computational work; and S. Wilder for help in the preparation of this manuscript. This study was supported by a National Science Foundation grant to M.G.R., an NIH grant to B.A.F., a Robert and Richard Rizzo Memorial Fund grant and an NIH grant to N.L.L., a Lucille P. Markey Foundation award to Purdue University, and a Purdue University reinvestment program grant. T.D. was supported by a European Molecular Biology Organization long-term postdoctoral fellowship for some of the time of this work.

Correspondence and requests for materials should be addressed to M.G.R. (e-mail: mgr@indiana.bio.purdue.edu). The coordinates and structure factors have been deposited with the Brookhaven Protein Data Bank, with accession numbers 1AL0 and 1AL05F, respectively.